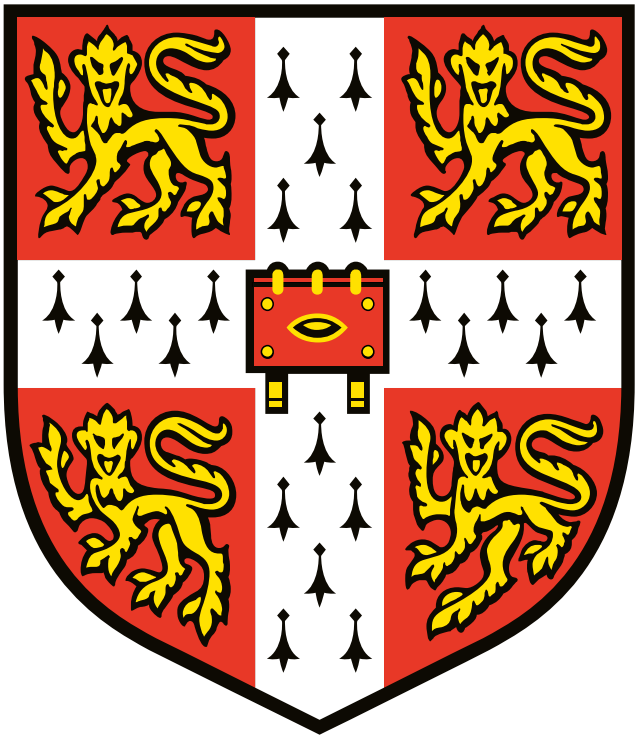
Investigating mechanisms of genome folding by single-cell Hi-C



Yang Cao

Selwyn College

Department of Biochemistry

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

June 2019

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Investigating mechanisms of genome folding by single-cell Hi-C

Yang Cao

Summary

In recent years researchers have begun to reveal the hierarchy of mammalian DNA folding from the 10 nm fibre to the intact whole genome. One of the key approaches used is the single-cell Hi-C.

My lab and I developed a single-cell Hi-C protocol that combines imaging and Hi-C processing on the same cell, and I tried to improve the sequencing library processing steps using a commercial Tn5 transposase. My work shows that without significant optimisation the commercial transposase is not compatible with our protocol.

Using the protocol, we successfully processed haploid mouse embryonic stem cells (mESCs) and calculated 3D structures of their entire genomes. The structures show several genome-wide features that are highly conserved across all mESCs, including: the genome shape is an ellipsoid; the chromosomes arrange in a Rabl configuration; the A and B compartments form a bowl-like structure; and active enhancers, promoters along with gene expression cluster together in the A compartment. In contrast, relatively local features were found to significantly vary from cell to cell, including the folding of: individual chromosomes, topological-associated domains and loops.

We also investigated mESCs in early stages of differentiation using the same single-cell Hi-C protocol. Preliminary analysis of chromosome and genome structures at 24 and 48 hour post differentiation reveals that certain features vary greatly. In particular, the genome shape of cells after 24 hour differentiation is relatively flat, whilst genomes after 48 hour differentiation are both ellipsoid and flat. Interestingly, the shapes of chromosomes from cells expressing the pluripotent marker *Rex1* after 24 hour differentiation are similar to the shapes of chromosomes in ES cells; whereas chromosome shapes from cells both after 24 hour differentiation with low *Rex1* expression and after 48 hour differentiation are even more varied.

Acknowledgement

First, I thank Prof Ernest Laue for giving me the opportunity to undertake this research project as a PhD student and for all his supervising throughout the past four years.

I thank all members of the Laue group for everyone's contribution to the project, with particular appreciation to Dr David Lando for his constant guidance on the wet lab and Dr Tim Stevens for his generous support for the computational work.

I must also thank all the collaborators for their help on the project.

I thank my college, Selwyn, and the University of Cambridge for a position to fulfil my PhD and a comfortable place to live.

Finally, I appreciate the enormous support of my family for my PhD and all their love throughout my life.

Table of contents

List of figures

List of tables

List of abbreviations

2i	two small molecule kinase inhibitors
3C	chromosome conformation capture
3D	three-dimensional
4C	circular 3C or 3C on chip
5C	3C carbon copy
A (amino acid)	alanine
A (base)	adenine
aa	amino acid
ATM	amplicon tagment mix
bp	base pair
C (amino acid)	cysteine
C (base)	cytosine
CENPA	centromere protein A
ChIP	chromatin immunoprecipitation
ChIP-Seq	chromatin immunoprecipitation followed by sequencing
cLAD	constitutive lamina-associated domain
CT	chromosome territory
CTCF	CCCTC-binding factor
DNA	deoxyribonucleic acid
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
E	glutamic acid
ESC	embryonic stem cell
FACS	fluorescence-activated cell sorting
FISH	fluorescence in situ hybridization

FU	fluorescent unit
G (amino acid)	glycine
G (base)	guanine
G_0-phase	gap 0 phase
G_1-phase	gap 1 phase
G_2-phase	gap 2 phase
GFP	green fluorescent protein
Gsk3	glycogen synthase kinase-3
h	hour
hESC	human embryonic stem cell
HiFi	high fidelity
I	moment of inertia
IE	inner end
Inh	inhibitor
iRFP	near-infrared fluorescent protein tagged
IS50L	insertion sequence 50 left
IS50R	insertion sequence 50 right
K	lysine
kb	kilobase
L	leucine
LAD	lamina-associated domain
LIF	cytokine leukemia inhibitory factor
LMP	low melting point
Mb	megabase
Mek	mitogen-activated protein kinase kinase
mESC	mouse embryonic stem cell
Mi-2	chromodomain-helicase-DNA-binding protein 3 or 4
min	minute

M-phase	mitotic phase
mRNA	messenger RNA
N	chromosome ploidy
NGS	next generation DNA sequencing
NPM	Nextera PCR mix
NT	neutralize tagment buffer
NuRD	nucleosome remodelling deacetylase
OE	outer end
P	proline
PCR	polymer chain reaction
poly(A)	polyadenylated
poly(T)	polythymidylated
RE1	restriction enzyme 1
RE2	restriction enzyme 2
RMSD	root-mean-square deviation
RNA-Seq	RNA sequencing
rRNA	ribosomal RNA
RT	room temperature
s	second
S	serine
S-phase	synthesis phase
STAT3	Signal transducer and activator of transcription 3
T	thymine
TAD	topological associated domain
TD	tagment DNA buffer
Tnp	transposase
W	tryptophan

1. Introduction

All cellular organisms and many viruses use deoxyribonucleic acid (DNA) to carry genetic information for life and reproduction. Such genomic information is primarily stored as codes of four types of bases, adenine (A), guanine (G), cytosine (C) and thymine (T), but is substantially more complicated than only a series of letters. Almost all cells in a multicellular organism contain the same DNA sequence which contains all the genetic information to encode all the cell types for that specific organism. A key question concerning modern day biology is how this DNA code is read by different cell types to confer their own function and phenotype. Increasingly overwhelming research suggests that the three-dimensional (3D) organization of the DNA sequence may play an important role in this process (for reviews see Ref[1,2]).

1.1. 3-Dimensional organisation of the genome

The total number of bases in a single mammalian genome is in the order of billions. These bases constitute tens of chromosomes which are linear DNA molecules with a total length of a few metres, but almost all these DNA molecules need to fit into a single cell nucleus which has an average diameter of only 6 μm. A key question is how the long DNA polymer fit into such a small 3D space. With newly developed technologies such as chromosome conformation capture (3C) and its derivatives, recent researchers have established a basic framework of DNA structural hierarchy (Figure 1.1.1).

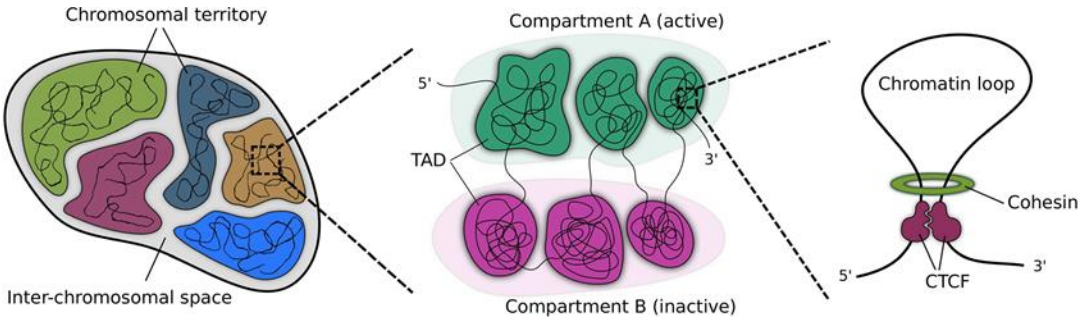


Figure 1.1.1 Basic DNA structural hierarchy

A genome (left) is compartmentalised into chromosomal territories, A/B compartments, topological associated domains (TADs) and chromatin loops (right). *This figure is reproduced from Szalaj P. et al.[1]

1.1.1. The building blocks – DNA double helix and the 10 nm fibre

The finding that the DNA polymer exists as a DNA double helix structure by X-ray crystallography in 1953[3] was a revolutionary discovery. This primary order of DNA structure helped form a solid basis for all the following research on DNA organisations and function. Until the late 1990s, the secondary order of DNA structure was determined to be the 10 nm fibre, a bead-on-a-string structure containing repeating nucleosome subunits[4]. Each nucleosome wraps approximately 146 base pairs (bp) of DNA, by means of a group of chaperone proteins called histones[4]. A single nucleosome normally contains one complex of histones from 5 different families, H1, H2A, H2B, H3 and H4. Four distinct homodimers of each of the core histones H2A, H2B, H3 and H4 constitute the nucleosome particle for DNA to wrap around. The linker histone H1 further locks the DNA by binding to the entry and exit sites. In addition, about 50 bp of DNA on average link two adjacent nucleosomes to form the bead-on-a-string structure. The varied length of linker DNA affects the flexibility of the structure, thus the tightness of DNA packing and DNA accessibility for molecules such as proteins to interact. DNA packing and accessibility are probably more significantly affected by post-translational modifications on histones. These histone modifications are known to change DNA-histone interactions,

nucleosome-nucleosome interactions and nucleosome-protein interactions[5]. Typical histone modifications include lysine acetylation, lysine/arginine methylation serine/threonine/tyrosine phosphorylation, lysine ubiquitination and arginine citrullination, but they are more commonly classified according to their functions, either activating or repressing genome activities such as gene expression. It is clear that the 10 nm fibre is necessary and sufficient to constitute chromosomes in most cell types and cell stages. The presence of the higher order DNA structure known as the 30 nm fibre has been largely debated over the past two decades and it existence is controversial [6–9].

1.1.2. Early studies of chromosomes

On the other hand, starting from the largest scale, chromosome morphology was first investigated by Walther Flemming using innovative microscopy and staining techniques in the late 19th century[10]. He found that chromosomes formed threadlike bodies that were able to move during mitosis[10]. These observations were later found to represent metaphase chromosome, the most condensed form of chromosome in eukaryotes. However, this form is predominantly found during the mitotic phase (M-phase) of the cell cycle. During interphase which represents the majority of cell life, chromosomes are in a more de-condensed configuration. There has been a large gap in our understanding between the 10 nm fibre and the metaphase chromosome. With newly developed techniques and approaches, researchers are now starting to slowly comprehend this gap.

1.1.3. Chromosome territories

At interphase, chromosomes are organized within the nucleus in discrete regions called chromosome territories (CTs). This concept was first proposed by Carl Rabl based on studies of epithelial cells of spotted salamander in 1885, and finally was

supported by some experimental evidence using different approaches such as UV (ultraviolet) laser damage labelling and DNA–DNA in situ hybridization in the late 20th century[11–15]. In the early 21st century, more convincing results using fluorescence in situ hybridization (FISH) microscopy and Chromosome Conformation Capture (3C) assays, further proved the existence of CTs[16,17]. The 3C assays probe spatially proximal information of two DNA loci. Briefly, this is carried out by crosslinking proximal DNA loci, fragmenting DNA, ligating the crosslinked fragments and identifying the loci based on sequence information of the ligated fragment (see Section 1.2 for detailed introduction). Intermingling between adjacent CTs was also found which suggests they are not completely insulated from each other. This potentially plays a role in nuclear activities like transcription[18]. There is evidence showing that CTs are not randomly arranged in the nucleus. Several studies have demonstrated a radial distribution of CTs, which is believed to be evolutionary conserved[19–22]. However this pattern was not found in nuclei of early blastomere stages of development. This suggests the formation of the radial distribution is associated with genome activation during early embryonic development[23]. This may contribute to the findings that the chromosome radial positions are correlated with local gene density [24–26]. Other factors such as geometric constraints and transcriptional activity were also shown to play roles in the genome radial arrangement[21,27]. So far, limited CT neighbourhood preferences were found only in specific cell types[28,29], and the proximity patterns established in a certain cell are not always conserved to the next interphase after a cell cycle[30,31]. Interestingly, a study on rod photoreceptor cells has shown that, although the CTs are still radially arranged in this specific type of terminally differentiated cells, the euchromatin and heterochromatin patterns within each CT are significantly different from the conventional zigzag pattern[32]. As a result the orientation and overall arrangements of the CTs have profoundly changed[33].

1.1.4. Genome compartments

For a specific cell type, within each CT and genome as a whole, DNA regions with similar properties tend to gather together to form mainly two types of clusters, termed A and B compartments. The A/B compartments were defined from the first genome wide 3C studies, called Hi-C, carried out by Lieberman-Aiden et al. in 2009[17]. Their genome wide contact matrices demonstrated a characteristic plaid pattern (Figure 1.1.4.1), where certain regions of a chromosome were found to interact and contact more preferentially with other regions located many Mb (megabase) away and vice versa. As well, the plaid pattern was observed in trans between chromosomes so that the regions of the same A/B type tended to correlate with one another while those of the opposite type anti-correlated over the entire genome (Figure 1.2.2.1 b compartments). Together these results also suggested that the correlated regions may tend to gather in 3D space within the nucleus. Bioinformatic analysis revealed that the A compartment is gene dense, structurally accessible and transcriptionally active. It also has enriched chromatin activating marks such as H3K36me3 and DNaseI (deoxyribonuclease I) hypersensitive sites[17]. In contrast the B compartment is not as gene rich, less accessible, late replicated and is correlated with lamina-associated domains (LADs)[34]. A more recent study using higher-resolution population Hi-C maps suggests the A/B compartments can be further partitioned into sub compartments, A1, A2 and B1-B4[35]. The A/B compartment profile is species and cell type specific. In a study comparing the profile in human embryonic stem cell (hESC) with the profiles in four hESC-derived lineages, 36% of the genome switched to the opposite compartment in at least one of the lineages[36]. Given such observed differences, similar to the CT arrangements, it is not clear yet how the A/B compartments change during cell differentiation.

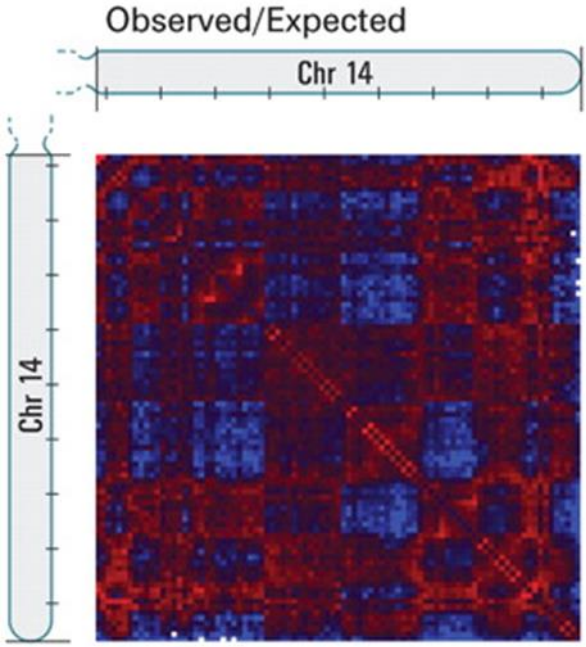


Figure 1.1.4.1 Plaid pattern of a Hi-C contact matrix

Hi-C contact matrix of chromosome 14 of human lymphoblastoid cell. The matrix is coloured according to the observed contact frequency between loci, either higher (red) or lower (blue) than expected based on their genomic distance. *This figure is amended from Lieberman-Aiden et al.[17]

1.1.5. Topological associated domains and CTCF/cohesin loops

Another level of DNA organisation that was observed in high-resolution population wide 5C or Hi-C studies restricts DNA to DNA interactions within certain regions of genomic sequence[37–39] (Figure 1.2.2.1 b domains). These regions of enriched contact frequencies where later called topological associated domains (TADs). Such restriction is thought to enhance the DNA interaction probability within each region. For example, the enhancers of these regions are much more likely to interact with promoters within the same TAD than promoter that are external. TADs are suggested to be formed as a consequence of looping events, where two boundaries restrict DNA interaction within a chromosomal section of approximately one Mb[40–42]. These looping events are achieved by dual binding of architectural factors CTCF

(CCCTC-binding factor, which binds to three regularly spaced repeats of DNA sequence "CCCTC") and cohesin[43]. The region between the two binding sites has been postulated to be extruded by a yet to be defined loop extrusion process and thus isolated from the rest of the chromosome[44]. During this process, a loop-extruding factor such as cohesin progressively extrudes DNA through its ring-like structure, until both ends interact with TAD boundary proteins such as CTCF and stop at these boundaries. This mechanism was suggested much later than the first definition of TADs, and is still being studied and discussed in the field. However it is becoming increasingly likely as evidence where the elimination of all loop domains were observed when the cohesin protein was degraded[45]. Also, the loop extrusion mechanism can also explain many properties of TADs, including some conflicts of earlier findings. First, TADs are conserved in animal evolution, where higher level of conservation occurs between closer relatives[37,39,46,47]. This is due to the conservation of CTCF binding sites especially between mammalian species[48]. Secondly, in a given organism, TADs were once known to be very compact and invariant across different cell types[36–38], and it occupies the majority of the genome. For example one suggested in mouse embryonic stem cells (mESCs), ~91% of the genome are partitioned into ~2200 TADs, with a median size of 880 kilobase (kb)[37]. The absence of certain TADs in part of the cell population was suggested based on DNA-FISH experiments[35], which argued against the invariance of TADs in a population. This in turn questioned the conservation of TADs across different cell types in an organism, which may be mistakenly inferred because of the conservation of potential CTCF/cohesin binding sites in their conserved genomes. The loop extrusion mechanism can also explain the formation of sub-TADs, smaller and more variable domains, by closer CTCF/cohesin binding sites[35,49]. Most CTCF sites in mammalian cells are potentially involved in mediating the sub-TADs, whereas only 15% are associated with TAD boundaries[43,50]. Generally speaking, at the moment this CTCF/cohesin mediated loop-extrusion mechanism for TAD formation still requires more evidences to ultimately prove the case.

1.1.6. Roles of proteins in DNA structure

In fact, DNA 3D organisations do not solely depend on DNA molecules themselves. As exemplified by histones, transcription factors, CTCF and cohesin, proteins play an important role in the formation of genome structures. Some protein-DNA interactions are shown to be mediated by chromatin remodellers. For example the NuRD (Nucleosome Remodelling Deacetylase) complex couples ATP-dependent chromatin remodelling with histone deacetylase activities. The NuRD/Mi-2 (Mi-2 analogous to Chromodomain-helicase-DNA-binding protein 3 or 4) complex could also recruit CTCF and cohesin (or other similar proteins) onto chromosomes[51–53]. The protein-DNA interaction network is still under intense research searching for mechanisms and functions involved in genome regulation.

1.2. Chromosome conformation capture

1.2.1. From 3C to Hi-C

Early studies on DNA structures were mostly based on microscopy techniques such as FISH. FISH has been widely used to investigate position information of genes or chromosomes by direct visualisation by specific DNA probes. However, the method is limited by probe specificity, low resolution and low throughput. In 2002, the introduction of 3C provided a more powerful way to study spatial proximity of two genomic regions by using intramolecular ligation and PCR (polymer chain reaction) to detect ligation frequency[54]. In 3C the experimentally detected two genomic sites that are close to each other in 3D space is called a “contact”. Its derivatives, 4C (circular 3C or 3C on chip), 5C (3C carbon copy) and Hi-C, further expand the 3C assay. Briefly, 3C probes a specific contact between two sites of the genome; 4C detects all contacts made by one specific site to the rest of the genome; 5C and Hi-C collect all contacts within a specific genomic region and the whole genome

respectively (Figure 1.2.1.1).

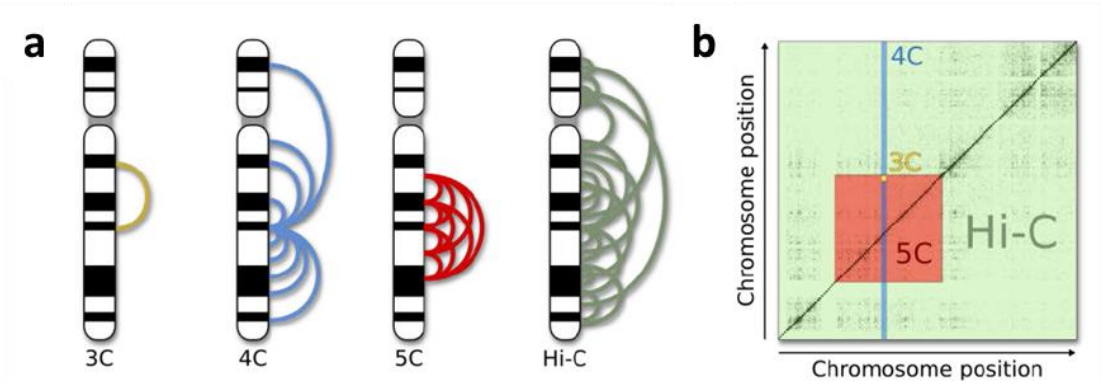


Figure 1.2.1.2 3C and its derivatives

a) Schematic of contact arrangements of 3C-based methods. The curves represent contacts between two sites of the chromosome detected by each method. In real cases, it is also possible to detect contacts between sites from different chromosomes. b) Contact map showing the number and the type of contacts detected by different methods, corresponding to the schematics in (a). The perpendicular axes are identical linear chromosome positions. Each dot in the map represents a contact between two positions of the chromosome/genome, which has a mirror image against the diagonal line. Example data of 3C are shown as the single yellow dot, 4C as the dots in the blue line, 5C as the dots in the red square area, and Hi-C as the dots of the entire map. *This figure is provided by Dr. Tim Stevens.

3C and its derivatives share basically the same experimental concept. First, the DNA sites that are proximal in 3D space are crosslinked by fixative agents such as formaldehyde. Next, chromatins are fragmented using a restriction enzyme, where the spatially proximal DNA fragments remain together via the crosslink. The crosslinked fragments are then ligated to form chimeric DNA molecules. Each ligated pair of DNA fragments forms a “contact”. These two fragments were proximal in space but are not necessarily proximal along the genomic sequence. After removing crosslinks, the contacts are identified using PCR or DNA sequencing (Figure 1.2.2.1 a).

1.2.2. Hi-C

Hi-C was first developed by Lieberman-Aiden et al. in 2009[17]. Compared with 3C, Hi-C has two key additional experimental steps to unbiasedly capture all-by-all contacts over the whole genome. After restriction, sticky ends are filled with biotinylated nucleotides. Ligation of the resultant blunt ends forms Hi-C junctions. Then after crosslink removal, DNA is sheared and only the biotinylated fragments are selected by means of biotin-binding streptavidin magnetic beads (Figure 1.2.2.1 a). These steps aim to improve the capture of valid DNA fragments containing a Hi-C junction in the sequencing library. Hence the sequencing data would be more informative as the invalid fragments do not waste sequencing capacity.

In recent years Hi-C has laid the basis for the study of whole-genome 3D organization at an unprecedented resolution[35,37,39]. At increasing levels of resolution, it is clear to see in Hi-C contact maps different levels of DNA 3D organisation at corresponding scales, such as chromosome territories, A/B compartments, TADs and loop structures (Figure 1.2.2.1 b). Various analyses on each level of organisation could then be done either using the Hi-C contact data themselves or in combination with data from other experimental approaches such as imaging, chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) and RNA sequencing (RNA-Seq).

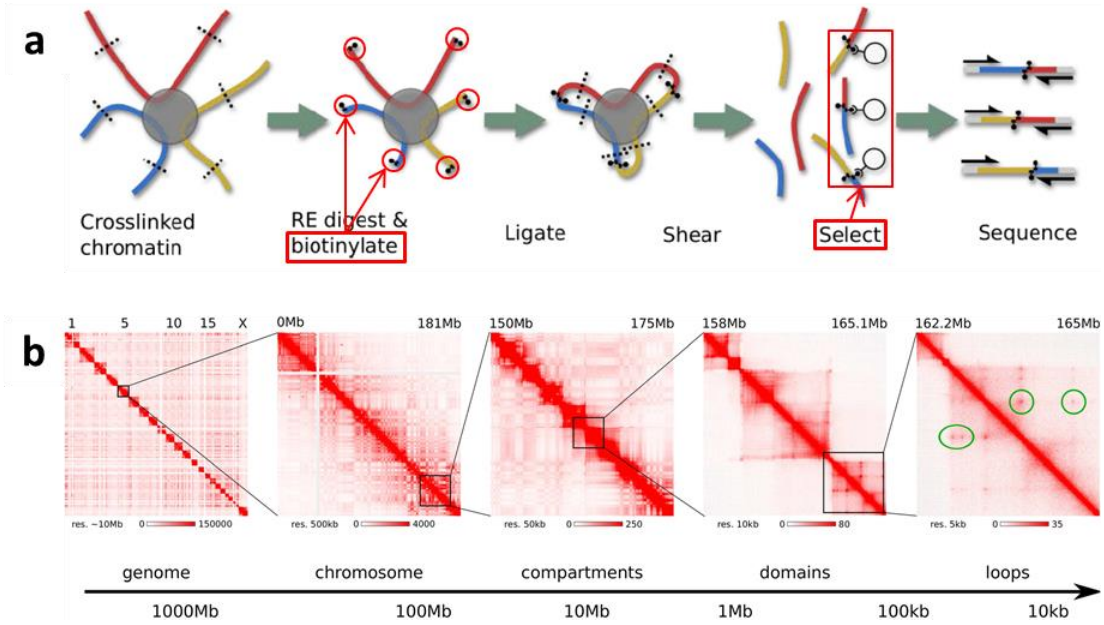


Figure 1.2.2.1 Hi-C

a) Simplified experimental workflow for Hi-C. The steps highlighted by red boxes and circles are Hi-C-specific steps that are not included in other 3C-based methods. *This figure is provided by Dr. Tim Stevens. b) Hi-C contact maps at different scales revealing different levels of DNA 3D organisation. The perpendicular axes are identical linear chromosome positions. Each dot in the map represents a contact between two positions of the chromosome/genome, which has a mirror image against the diagonal line. In general, the contact map of the whole genome shows dense contacts within the squares of each chromosome region along the diagonal line, representing chromosome territories. Dense squares are also visible at Mb-level maps, corresponding to TADs. Plaid patterns can be found in the maps of genome, chromosome and compartment levels, which refer to the A/B compartments. At high-resolution maps, individual peaks (highlighted by green ellipses) are clearly seen on the sides of squares, these represent loop structures. *This figure is adapted from Szalaj P. et al.[1] Heatmaps were created from the GM12787 in situ Hi-C dataset published by Rao et al. (2014) using Juicebox (Durand et al. 2016)[35,55].

Conventionally, Hi-C is carried out on population of cells. It collects contacts from the genomes of many cells and the results represent an average ensemble of all conformations. The information potential of a Hi-C dataset largely depends on the

amount of contact data or in general the resolution of the contact map. As well the resolution of a Hi-C contact map depends on the complexity of the Hi-C contacts identified from the sequencing data of the library. A dataset with a few billions Hi-C contacts can achieve a contact bin resolution as low as 1 kb[35]. Typically, a high-resolution Hi-C experiment requires millions of cells. More contacts or higher resolution could potentially be achieved by improving the reaction efficiencies, or by using a larger number of cells as starting material.

1.2.3. Single-cell Hi-C

In contrast to conventional so-called "bulk" or "population" Hi-C studies, where an average conformation from millions of cells is studied, newly developed single-cell Hi-C approaches allow the modelling of single chromosome structure revealing cell-to-cell variability[56]. This unique characteristic of being able to sample cell-to-cell variability is one of the main advantages of single-cell Hi-C over population Hi-C in studying DNA 3D organisation and function. It can investigate some structural features that could either be buried or incorrectly inferred due to conformation averaging in population Hi-C studies. For example, in the first single-cell Hi-C study differences between different cells were observed – in particular inter-chromosomal contacts formed cell-specific clusters in different single-cells, whereas it was not possible to see this feature in Hi-C studies carried out on populations of cells[56].

However, the low amount of starting material (2.5 pg for a single haploid mouse genome) is always a limiting factor in single-cell experiments, because any inefficiency in the various processing steps cannot be compensated for by increasing the amounts of starting material. In single-cell Hi-C experiments, inefficient steps lead to a lack of detection of valid fragments and thus a reduction in the resolution of modelled chromosome/genome structures, and the only way to better results is to improve reaction efficiency of every step in the single-cell Hi-C methodology. The other general challenge in current single-cell Hi-C studies is to get adequate good

quality datasets, where each dataset contains the information of an individual genome. Compare with population Hi-C, where the average conformation is already based on contact frequencies/probabilities of a cell population, single-cell Hi-C studies would always be questioned whether the limited sample size can represent the population. Obviously, more samples would be statistically better to demonstrate cell-to-cell similarity and variability.

1.3. Other methods complementary to single-cell Hi-C

A key characteristic of single-cell Hi-C data is that all the contacts are derived from a single genome that can be precisely positioned along the reference genome sequence. The datasets on their own are already very informative for DNA 3D organisations. In addition, they are good platforms to map data from other methods, such as ChIP and RNA-seq, which also possess sequence-positional information of the same genome. These correlative genomic approaches have great potential in studying the relationship between nuclear structure and function.

It should be noted that among the approaches mentioned above, some like ChIP-seq are currently not feasible to carry out at the single-cell level. Even for approaches like RNA-seq which can be carried out at a single-cell level, it still hasn't been shown that these experimental approaches are compatible with single-cell Hi-C carried out on the same cell. In other words, an applicable single-cell method should not influence either the genome structure itself for Hi-C or the single-cell Hi-C procedure. In any case, data from a population-based method does provide an initial opportunity for correlative analysis with single-cell Hi-C.

1.3.1. Imaging

In late 19th century, the first images of chromosomes was revealed by light microscopy and staining[10] (see Section 1.1.2). Then in the early 21st century, FISH

confirmed the presence of CTs by means of whole chromosome painting[16] (see Section 1.1.3). FISH could also probe loci of specific genes or DNA sequences, using fluorescently labelled complementary sequences. Innovative super resolution imaging techniques can image chromatin fibres at nanometre resolutions[57–59]. All these techniques may be done in vivo, with the potential in investigating dynamic information of chromatin structure. Imaging could potentially be compatible with single-cell Hi-C, because the imaging process may not affect genome structure or the feasibility of Hi-C reactions.

1.3.2. ChIP-Seq

Chromatin immunoprecipitation (ChIP) is a population-based technique used to study DNA-protein interactions. In brief, the method includes the following key steps. DNA and associated proteins are crosslinked together in-vivo inside the cell. Next the genomes are fragmented into small DNA pieces and DNA-protein complexes are then purified from other contents in the cells by targeting the protein of interest using an antibody. After crosslink removal, the associated DNA fragments are purified and their sequences are determined. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) combines ChIP with massively parallel DNA sequencing[60]. This allows precise mapping of the DNA sequences to its reference genome, which annotate global sequence-positional information to the protein of interest. As noted in the introduction of Section 1.3, unfortunately ChIP-Seq cannot be performed on the same cells that would be used for single-cell Hi-C. But the data of ChIP-Seq, especially from the same cell type, can be used for correlative analysis with single-cell Hi-C.

For our single-cell Hi-C experiments on mESCs or early differentiated cells from mESCs, it is particularly useful to map ChIP-Seq data of certain pluripotency transcription factors such as *Nanog* and *Klf4*, and data of typical histone modifications onto our single-cell datasets. This would give information about their functioning sites

and frequencies, allowing studies on relationship between genome structure and activity.

1.3.3. RNA-Seq

Typically, RNA sequencing (RNA-Seq) is a population-based technique used to reveal the transcriptome of a cell population. For genomic studies such as single-cell Hi-C, the nuclear fraction of messenger RNA (mRNA) is measured using this technique. In brief, the method includes the following steps. RNA is first isolated from tissue, with DNA degraded using DNase. Next, mRNA with 3' polyadenylated (poly(A)) tail is filtered by polythymidylated (poly(T)) oligomers. The abundant ribosomal RNA (rRNA) is then depleted by complementary oligomers. By reverse-transcribing the mRNA, a cDNA library is generated for massively parallel sequencing. The resultant transcriptome data could annotate genes in our single-cell Hi-C genome with their transcription activity. This would allow analysis of the relationship among gene position, genome structure and transcription. Similar to ChIP-Seq, the population-based and even single-cell RNA-Seq cannot be performed on the same cells that would be used for single-cell Hi-C. Data for correlative analysis need to be collected from different populations of the same cell type.

1.4. Cell cycle and early differentiation of mammalian ESC

1.4.1. Pluripotent embryonic stem cells

Embryonic stem cells (ESCs) are cells derived from the blastocyst stage of early mammalian embryos[61,62]. The most remarkable characteristic of ESCs is their ability to differentiate into any cell type of the species – so-called naive pluripotency[62]. As

stem cells, they are also able to self-renew indefinitely if they are kept under certain conditions in an undifferentiated state[63,64]. These properties bring ESCs great potentials for basic scientific research and biomedical uses. ESCs of model mammalian species such as mouse, or *Mus musculus* (Latin binomial name), are widely used to study mammalian biology, either at the ESC stage or as a source for specific differentiation pathways. Many of these studies lay the ground work for further studies in human cells, where human ESCs are used in research to study specific disorders and clinical therapies.

In my lab's single-cell Hi-C experiments, we used mouse ESCs (mESCs) to study mammalian genome architecture. There are many well established protocols for ESC culture, pluripotent proliferation, directed differentiation and control of these processes. So ESCs allows studies of different stages of mESC cell cycle and on changes when being induced into a specific differentiation pathway.

Identifying single-cell Hi-C contacts largely relies on mapping sequence reads to a reference mouse genome. Ideally one of the two contacting DNA fragments should be mapped to a unique position in the genome. For the experiments described in this thesis I used haploid mESCs instead of diploid cells. The reason for using haploid cells was that haploid contained only one copy of each chromosome and mapping to the correct chromosome was relatively straightforward. If diploid cells were used, the homologous chromosomes would almost have the same sequences, except for single nucleotide polymorphisms. Mapping Hi-C DNA fragments to the correct homologue would therefore be more difficult, increasing the number of ambiguous contacts and making structure calculations more unreliable.

A consistent method of generating an expanded population of haploid mESCs has been developed by Leeb et al. in 2011[65]. It derives haploid mESCs from unfertilised female mouse oocytes. During the expansion process, as haploid cells tend to become diploid, the haploid population can be successively enriched by staining DNA and flow sorting cells with haploid DNA content. The resultant population have been shown to have genetic integrity of a haploid mouse genome. Their genome-wide expression profile and cell morphology are both similar to diploid mESCs.

1.4.2. ESC self-renewal and cell cycle

If ESCs are trapped in the pluripotent state by certain conditions and cannot differentiate, they are capable to indefinitely proliferating through the cell cycle (as shown in Figure 1.4.2.1 for haploid cells). The cell cycle has four main phases, G_1 (gap 1), S (synthesis), G_2 (gap 2) and M (mitosis) phase, where G_1, S and G_2 phases together are called interphase. In G1 phase, cells grow and prepare for DNA synthesis. Chromosomes stay in normal ploidy, that is, 1N (one copy of chromosome) ploidy for haploid mouse/human cells and 2N ploidy for normal diploid mouse/human cells with a pair of homologous chromosomes. In S phase, DNA synthesis generates a copy for each chromosome, called sister chromosome, while ploidy gradually doubles. This means for haploid cells ploidy is changing from 1N to 2N, and for diploid cells it is changing from 2N to 4N. When DNA synthesis completes, cells continue growing and prepare for mitosis in G_2 phase and ploidy stays the same. Finally, in M phase a mother cell divides into two daughter cells, each sister chromatid ends up in one of the daughter cells and ploidy halves (1N for haploid cells and 2N for diploid cells) for each daughter cell. For certain reasons such as nutrient depletion, cells in G_1 phase may enter G_0 phase for resting, and may return back to G_1 phase if they are allowed to grow again.

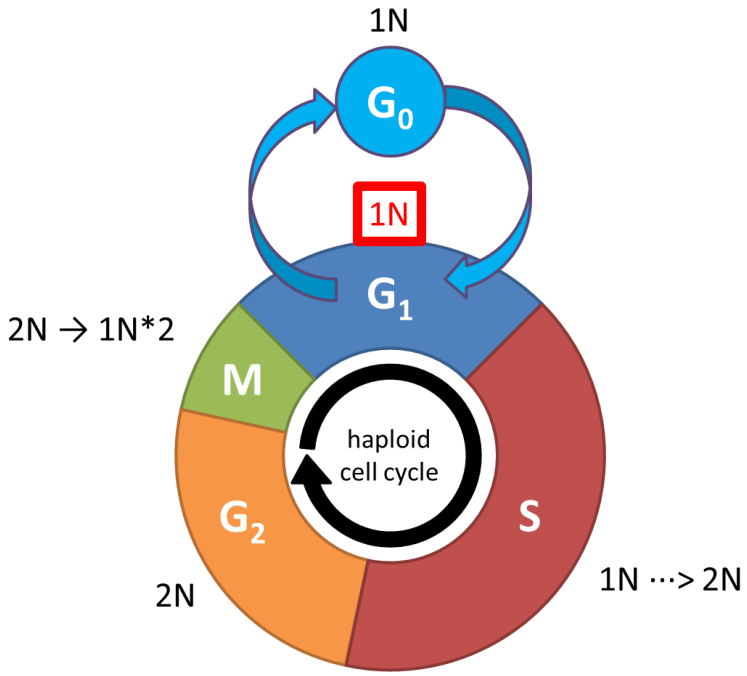


Figure 1.4.2.1 Schematic of ESC cell cycle

The cell cycle of a haploid cell is shown and coloured according to different phases. The ploidy (N) of each phase is labelled at the periphery.

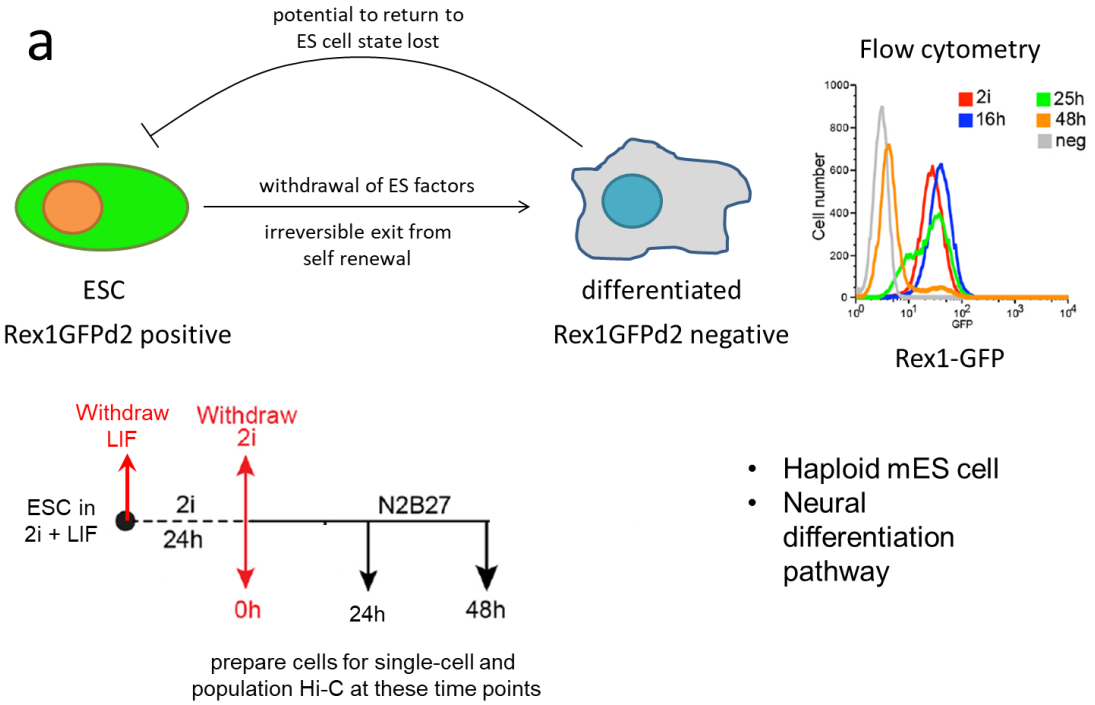
As discussed in the previous section (Section 1.4.1), my lab and I decided to use haploid mESCs to avoid ambiguous mapping of Hi-C contacts. For the same reason, we aimed to study genome architecture in the phase with only one copy of each chromosome (1N), or the G_1 phase for haploid cells (as indicated in Figure 1.4.2.1). This assumes the cells are continuously self-renewing, not entering the resting G_0 phase. We developed a strategy using Fluorescence-activated cell sorting (FACS) to sort G_1 cells from cells in other cell cycle phases[66]. We also used imaging of the centromere protein A (CENP-A) to check chromosome number and thus whether the G_1 sorting is successful for each single cell[66] (see Section 1.3.2 and method Chapter 7). It is also interesting to investigate genome conformation changes during the cell cycle, in other words, how chromosomes fold into chromatids and unfold back into their territories. Some preliminary studies have been carried out by Nagano et al.[67].

1.4.3. ESC early differentiation

Certain environmental conditions can stop ESCs from self-renewing and induce differentiation. As mentioned in Section 1.4.1, pluripotent ESCs are capable to differentiate into any cell type. Directed differentiation into specific cell pathways is also possible by applying specific conditions to the ESCs (for review see Terryn et al.[68]).

Controlled early differentiation of ESCs can be achieved by growing them in 2i/LIF conditions. The combination of 2i (two small molecule kinase inhibitors) and LIF (cytokine leukemia inhibitory factor) conditions can derive pluripotent ESCs and maintain the pluripotency[63]. In particular, the two inhibitors of 2i are PD0325901 and CHIR99021, which inhibit mitogen-activated protein kinase kinase (Mek) and glycogen synthase kinase-3 (Gsk3) respectively. These two kinases are known to be involved in triggering differentiation. LIF can activate STAT3 (Signal transducer and activator of transcription 3), which is required for ESC self-renewal[69]. During the process, pluripotent mESCs are grown in a neural media (N2B27) containing both 2i and LIF. Upon LIF removal and with 2i only, the cells stopped proliferating; and upon 2i depletion, the cells started to differentiate into the neural pathway because of the media[70]. The strategy can monitor mESC early differentiation based on the mRNA changes of a pluripotent marker gene, *Rex1* (Figure 1.4.3.1 a). In undifferentiated mESCs, *Rex1* is highly expressed with upregulation by pluripotent factors such as *Nanog* and *Sox2*[71]. But it is severely and abruptly downregulated during early differentiation[72]. The cells after 24 hours' differentiation form a mixed population expressing either high or low level of *Rex1*, whereas after 48 hours almost all cells show a low expressing level[70]. Differentiated cells with low level of *Rex1* lose the potential to self-renew or to return to the ESC state (Figure 1.4.3.1 a)[70]. Samples for single-cell Hi-C can be collected at certain time points to generate a view of genome architecture changes during this differentiation process (Figure 1.4.3.1 a). The strategy was also used in RNA-Seq experiments which monitored changes of gene expression level[70] (Figure 1.4.3.1 b). This provided insights into the relationship

between changes in genome structure and changes in transcription and function during these early differentiation steps.



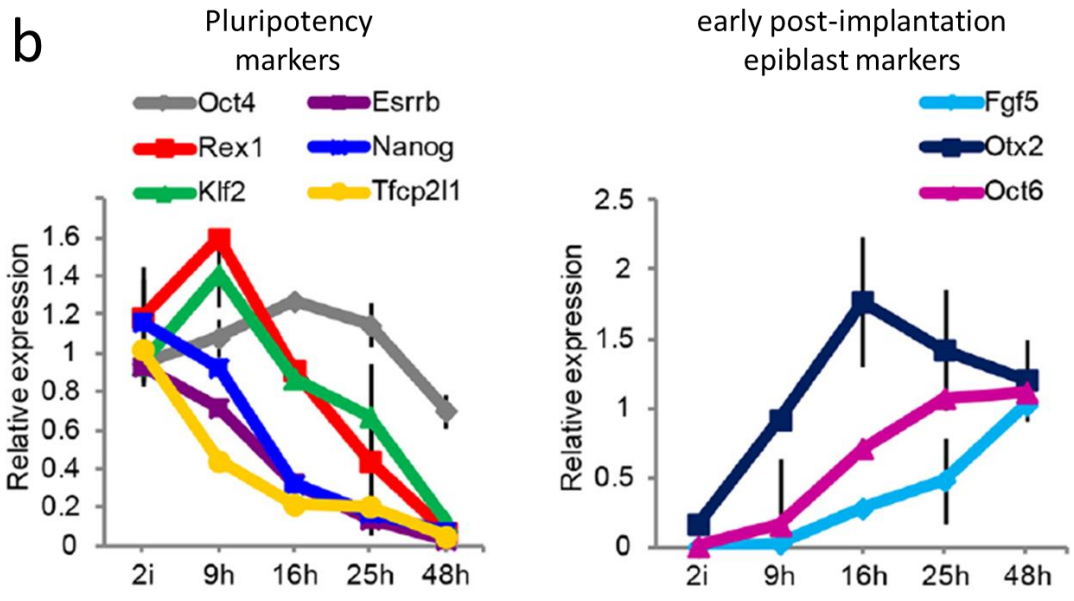


Figure 1.4.3.1 Cell preparation and complementary gene expression data for single-cell Hi-C studies on early differentiated mESCs

a) Experimental design of preparing early differentiated cells for single-cell Hi-C. Flow cytometry of green fluorescent protein tagged (GFP-tagged) *Rex1* was used to monitor the

differentiation process. mESCs remain in the pluripotent ground state in N2B27 media containing 2i and LIF. mESCs exit ground state by removing LIF and incubating with 2i for 24 hours. Differentiation process was induced by removing 2i and culturing cells in N2B27 media. Cells for Hi-C were prepared at the indicated time points. b) Relative gene expression level changes of selected pluripotency and early differentiation markers during mESC early differentiation. *Figures are amended from Kalkan T. et al.[70]

1.5. Tn5 transposase

A transposable element, or a transposon, is a DNA sequence that can be moved to other positions of a genome. Class I transposons are also called retrotransposons, which function by a "copy and paste" mechanism via intermediate RNA and reverse transcription. Class II transposons or DNA transposons, however, are excised from their original positions and inserted into other positions. This conservative "cut and paste" mechanism, transposition (Figure 1.5.1 a), is performed by a group of enzymes called transposases, which are encoded by the transposons themselves. Tn5 transposon is a well-studied model system of transposition. The wild-type Tn5 transposon found in bacteria contains two nearly identical IS50 elements (IS50L and IS50R, insertion sequence 50 left and right), each enclosed by two 19-bp DNA end sequences (IE and OE, inner end and outer ends), and an embedded region between the elements (Figure 1.5.1 b). The 476 aa (amino acids) Tn5 transposase is encoded from IS50R. During transposition, two transposases each bind to a specific DNA end sequence (see Figure 1.5.1 a for all of these steps). Then the transposase-DNA complexes dimerise to form a synaptic complex and a trans-acting catalytic site for cleavage (see Figure 1.5.1 c for dimer structure). The cleavage involves formation and re-cleavage of a hairpin structure by the 3' and 5' strand at the resultant blunt end, which allows the cleavage of both ends of the transposon without major reorientation of the only trans-acting catalytic site. The complex containing the cleaved transposon then binds to the target DNA, where for the wild-type Tn5 transposase a specific 9-bp

sequence is only required for insertion but not binding[73].

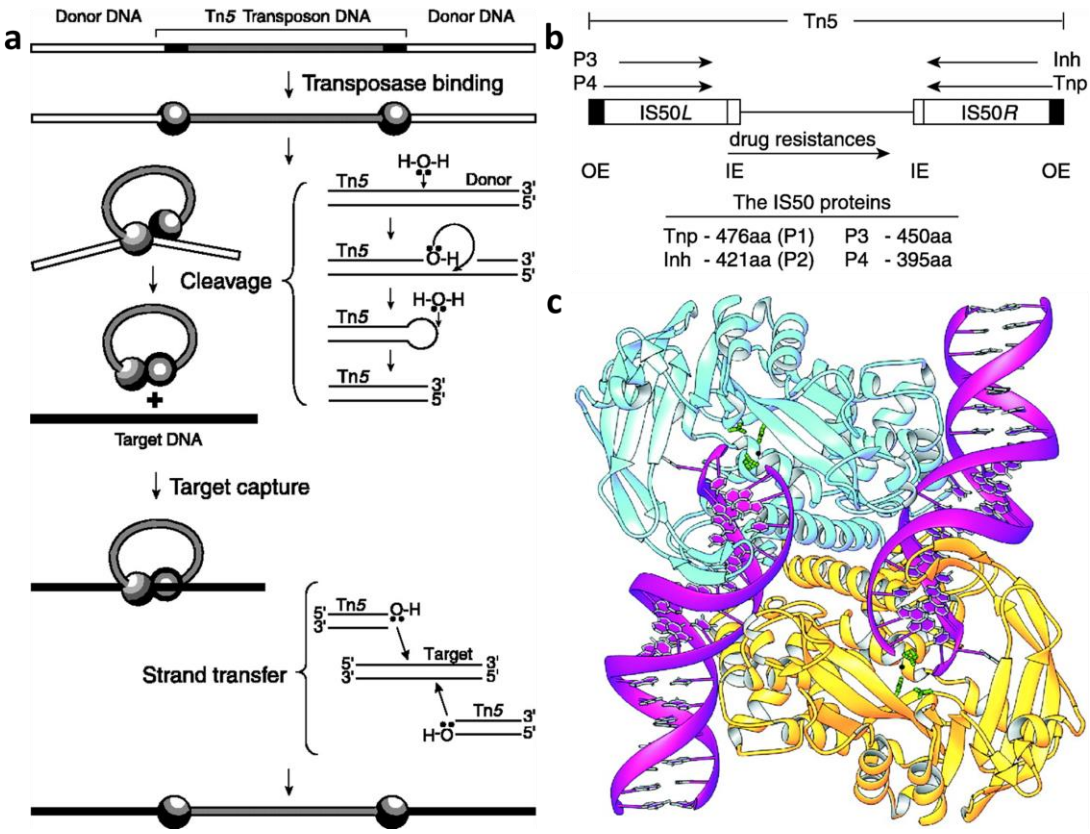


Figure 1.5.1 Tn5 transposase.

a) Schematic of Tn5 transposition. *This figure is amended from Reznikoff et al. and Davies et al.[74,75]. b) Schematic of bacterial Tn5 transposon. The two IS50 elements are enclosed by two 19 bp DNA, OE and IE. IS50R encodes the Tn5 transposase (Tnp) and its inhibitor (Inh), IS50L encodes the inactive C-terminal truncated version of Tnp (P3) and Inh (P4) and the embedded region encodes for antibiotic resistance. The table shows the sizes in number of amino acids of the encoded proteins. *This figure is amended from Goryshin et al.[76]. c) Ribbon representation of the structure of Tn5 transposase-DNA dimer studied by X-ray crystallography. The two subunits of transposases are coloured in yellow and pale blue. The two 20 bp DNA molecules are coloured in purple. *This figure is amended from Davies et al.[75].

The wild-type transposases are usually structurally closed and inactive to maintain genome stability. But their unique function and molecular mechanism lead to great

potential as genetic tools for experiments such as targeted insertion or deletion (for review see Ref[77]). These experiments are carried out by replacing the wild-type transposon by recombinant transposons. As the model system, the hyperactive version of Tn5 transposase is well studied and several hyperactive mutations has been identified, including W450C, W450S, L372P, E54K, E344A, P242A, P242G and E110K (one-letter aa code: W-tryptophan, C- cysteine, S-serine, L-leucine, P-proline, E-glutamic acid, K-lysine, A-alanine and G-glycine)[78–82]. For example the L372P mutation opens the conformation and activates Tn5. In addition, the wild-type 19-bp end sequences (IE and OE) in the transposon are relatively inactive for Tn5 transposition. To activate transposition, they can be replaced by hyperactive mosaic ends (ME) in recombinant transposons[83,84]. The nonspecific binding characteristic of Tn5 transposase provides additional possibilities in DNA library preparation experiments. Syed et al. developed the reaction to simultaneously fragment DNA and ligate adapters to both ends using Tn5 transposase and ME-integrated oligonucleotide complexes called transpososomes[85,86]. The reaction is called tagmentation and it was further developed using a hyperactive variant of Tn5 transposase to a 5-min reaction that can target as little as 10 pg DNA[87]. Combined with the next generation DNA sequencing (NGS), Tn5 transpososomes has become a powerful library preparation tool for DNA sequencing projects such as single-cell Hi-C[88,89]. The most recent commercial version of this tool is the Nextera XT DNA Library Preparation Kit[90].

1.6. Key goals and project aims

1.6.1. To develop an improved single-cell Hi-C protocol

When I started my PhD, single-cell Hi-C was an emerging technique and my lab was involved in the development of the first single cell Hi-C protocol[56]. However, this initial protocol identified only around 10,000 contacts per cell, not enough to calculate chromosome and genome structures below 500 kb resolution. Given the theoretical

maximum of distinct mappable fragment-end pairs per single cell (1,201,870 for BglII as RE1, restriction enzyme 1), the recovery rate was only up to 2.5%. To address this, the main goal of my PhD project was to develop the next generation of the single-cell Hi-C method, with the aim of improving both the efficiencies of some key reaction during "wet" processing and the reliabilities of sequencing data analysis steps during the "dry" processing. This included developing and testing a new method for library preparation that used Tn5 transposase. In addition, my lab and I also aim to develop ways to combine single-cell Hi-C with other complementary methods (see Section 1.3), to validate the single-cell Hi-C method itself and better understand genome structure and function.

1.6.2. To calculate and study whole genome architecture of single mESC at 100 kb resolution

As discussed in Section 1.4.1 and 1.4.2, mESC is an important cell type that carries fundamental information about mammalian genome biology. Better understanding genome structure and its relationship to genome function in mESCs would potentially provide clues for genome structures of other mammalian or even non-mammalian cell types. As well the results would also help to understand at the single cell level the observations and findings made from population based Hi-C experiments and related work. For example, to what extent do TADs, loops and compartments exist in single cells, and if so how are they organised at a single genome level. Studying genome structure in single cells would therefore help investigate any information lost due to conformation averaging in population studies. Finally, all of these aims would be better undertaken using calculated single-genome structures at higher resolution, which is in-line with the previous aim outlined in Section 1.6.1.

1.6.3. To apply single cell genome Hi-C to study genome architecture changes during early mESC differentiation

After successful method development and getting enough results for mESCs, studies on consecutive time points along the early differentiation process would draw a clear view of the changes in genome structure from mESC. These changes could also be correlated with changes in biological activities such as gene expression. Comprehensive understanding of both changes due to differentiation would possibly provide insights in relationship between genome structure and function, and clues about genome dynamics.

2. Experimental design

Other contents of this thesis may refer to specific parts or steps of our combined imaging single-cell Hi-C protocol. To help correlate the methods, results and rationale, and to avoid confusion, this chapter introduces the experimental designs.

2.1. Single-cell Hi-C experimental design

The combined imaging and single-cell Hi-C workflow is shown in Figure 2.1.1. The method was developed by my lab and I during the 4 years of my PhD and here I only introduce the current developed version.

I normally aimed to process 20 single cells per experiment. In some experiments fluorescent labelling of proteins would be imaged (see Section 1.3.2) while in some of the other experiments samples were induced to differentiate so that early differentiation studies could be carried out (see Section 1.4.3 and Chapter 5). My colleagues S.B. (Dr. Srinjan Basu) and D.L. (Dr. David Lando) prepared haploid mouse embryonic stem cells (mESCs) for me. The cells were haploid sorted to avoid becoming diploid, by fluorescence-activated cell sorting (FACS) every 1-2 weeks during cell culture[91]. A single cell suspension was fixed with formaldehyde, and DNA sites that were close together in space were potentially crosslinked by formaldehyde molecules (Figure 2.1.1 step 1). The fixed nuclei were isolated and then permeablized to facilitate the entry of reagents for Hi-C reactions (Figure 2.1.1 step 1). Single nuclei were then isolated and sorted into individual wells of 384-well plates using FACS, and then covered with NEBuffer 3 to avoid drying out during imaging (Figure 2.1.1 step 2).

Then S.B., D.L. and I carried out imaging cooperatively. To obtain 20 imaged samples for processing we normally had to scan 40 wells to find 20 single cell nuclei at this stage. We did this to confirm that the sample: was not a multi-cell sample; was a haploid nucleus in G1 phase by imaging centromere protein CENP-A and confirming

the number of centromeres was correct; and looked like a normal healthy nuclei by white light imaging (Figure 2.1.1 step 3). Images of mEos3.2 labelled CENP-A protein were also used for structure validation later (see Sections 1.4.3 and 4.2).

The following Hi-C reaction processing was carried out by D.L. and S.B. for mESCs and by me for differentiated cells. After finding 20 cell nuclei to process, each selected nuclei was covered with an agarose pad made from low melting point (LMP) agarose (Figure 2.1.1 step 4). This pad trapped and immobilized the nuclei, and allowed parallel Hi-C processing by enabling reaction solution exchange without disturbing or losing the nuclei. Nuclei were then Hi-C processed individually in parallel by the following reactions (Figure 2.1.1 steps 5 – 7). Crosslinked DNAs were fragmented by a restriction enzyme (referred as restriction enzyme 1 or RE1, we used MboI in most experiments and BglII in some of the early experiments), leaving sticky ends at both ends of the DNA fragments. This reaction cut the genome at various positions while the crosslinked DNA fragments remained associated. The digested sticky ends were filled-in with nucleotide mix containing biotinylated adenine nucleotides generating blunt ends (Figure 2.1.1 step 6). The blunt ends were then ligated and formed Hi-C junctions (Figure 2.1.1 step 7), before the crosslinks were removed (Figure 2.1.1 step 8).

For the sequencing library preparation steps, D.L. processed most mESCs using the AluI-A-tailing method; I processed some mESCs by the AluI-A-tailing method, all mESCs by the transposase method and all differentiated cells by the AluI-A-tailing method (no differentiated cells were processed by the transposase method) (Figure 2.1.1 steps 9 – 13). The two methods will be compared in more detail in the next section. In brief, the LMP agarose pad was melted, either straight after the crosslink removal or by reheating if plates were once stored in fridge. The molten agarose solution was kept melted while the biotin-labelled Hi-C junctions were bound to streptavidin-coated magnetic beads (Figure 2.1.1 step 9). Magnetically separating and then washing the beads removed the agarose. Bound Hi-C fragments were then digested with an AluI blunt end restriction enzyme (referred to as restriction enzyme 2 or RE2) (Figure 2.1.1 step 10). After adding a 3’-deoxyadenine nucleotide, adapters

containing barcodes and PCR primer sequences were ligated to the Hi-C fragments. (Figure 2.1.1 step 11). The transposase method also employed the AluI trimming step (Figure 2.1.1 step 10a), before tagmentation added adapter sequences in one reaction without the need to A-tail the fragments. (Figure 2.1.1 step 11a). For PCR amplification of Hi-C junctions the samples were transferred from the 384-well plate into individual PCR tubes (Figure 2.1.1 step 12). We found the PCR amplification was more consistent when carried out in tubes than 384-well plates. Then the libraries were amplified (Figure 2.1.1 step 13), purified (Figure 2.1.1 step 14) and analysed to see fragment distributions and yields. Libraries with good fragment distribution and yield were selected and pooled for sequencing.

The number of selected libraries depended on the efficiencies of all reactions in the protocol, i.e. steps 5 – 14 in Figure 2.1, and varied in different experiments. Libraries from different experiments might also be pooled together as long as they were from the same library preparation method (AluI-A-tailing or transposase), with different sequencing indexes/barcodes, and within the sequencing capacity. Then the pooled library was size selected for fragments in the range of 300 – 700 base pairs (bp), to allow both effective sequencing and unique genome mapping (Figure 2.1.1 step 15). After checking the size selection was successful, the library was sent for high-throughput sequencing (Figure 2.1.1 step 16). In certain cases, I checked the library sequence qualities by relatively low throughput Illumina Miseq sequencing first. Then the ones with promising sequence qualities were pooled and size selected again for deeper sequencing using Illumina HiSeq 4000.

After sequencing the reads were demultiplexed into individual sample libraries, according to their indexes/barcodes. Reads for each library were then processed and analysed and Hi-C contacts were identified using NucProcess, a Python software developed by our group (Figure 2.1.1 step 17). If the contact profile was good, a bead-on-a-string model of genome structure was calculated based on restraints generated from the contacts.

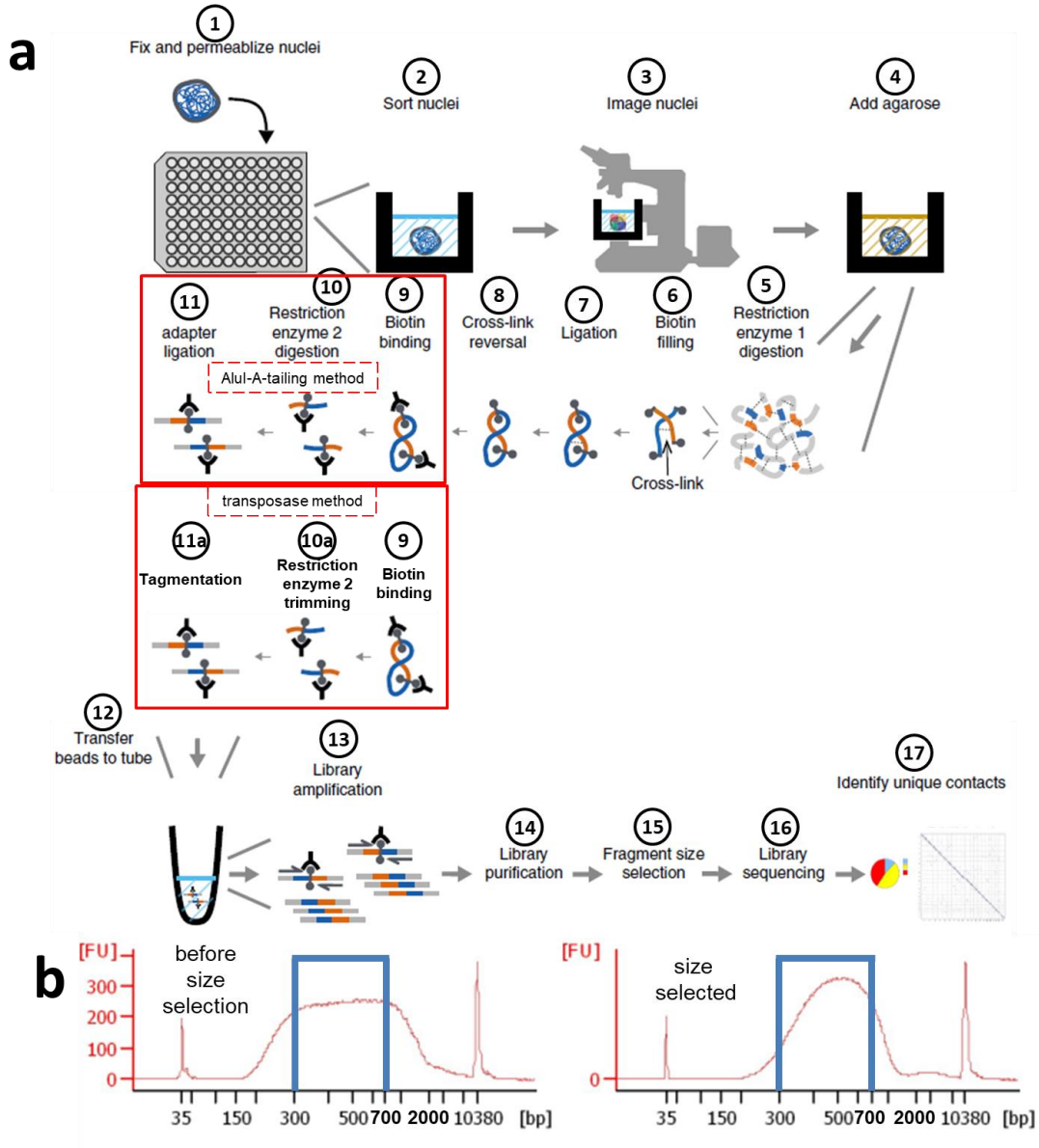


Figure 2.1.1 Schematic workflow of the single-cell Hi-C protocol

a) The workflow follows each step in sequence from 1 to 17, as indicated by grey arrows. Steps 9 – 13 can be carried out in either the AluI-A-tailing method or the transposase method, where the key differences are shown in red boxes. b) In step 15, fragment distributions of pooled library before (top) and after size selection (bottom) are compared by their electropherograms. The blue boxes indicate the optimal range of fragment size between 300 and 700 bp. *This figure is amended from Lando et al.[92].

2.2. Comparison of sequencing library preparation methods for single-cell Hi-C

As discussed in Section 1.6.1, the original single-cell Hi-C method used in the first single-cell Hi-C experiment by Nagano et al.[56] has a low efficiency identifying Hi-C contacts. One of the main causes was thought to be the AluI-A-tailing library preparation reactions. During the process, the input DNA is first fragmented by the restriction enzyme, AluI, leaving blunt ends, then tagged with a 3'-deoxyadenine (A) nucleotide and finally ligated with PCR primers using the A-overhang. These three reactions take more than 3 hours in total and all of them are potentially inefficient. In contrast, as discussed in Section 1.5 the Tn5 transposase can combine the three into one 5-min reaction, with high efficiency and the ability to target picogram amount of DNA[87]. This provides a possible alternative to the original library processing steps for single-cell Hi-C, with potential improvements on reaction efficiency.

As mentioned in Section 1.6.1, one of the main goals of my PhD was to integrate the Tn5 transposase system into single-cell Hi-C. The relative results will be discussed in Chapter 3. I was also involved in optimising the original method, aiming to improve the capture of Hi-C contacts[66]. This section compares the workflows of the two methods in details. Only the latest developed versions of both workflows are described.

For the developed transposase method with trimming incorporated, the bead binding, LMP agarose removal and AluI restriction steps were the same as the AluI-A-tailing method, except lower amount of AluI was used in restriction (Figure 2.2.1 steps 1 and 2, compare left with right). After crosslink removal and while keeping the LMP agarose melted, biotinylated fragments were bound to streptavidin-coated magnetic beads. This was carried out by 1 h incubation at 37℃ with bead slurry. After bead binding while keeping the plate at 37℃, the magnetic beads were separated from the solution by a magnetic stand, and liquid LMP agarose was removed by pipetting. The beads were washed for a few times to purify the bound biotinylated DNA, exchange

the buffer and dilute any residual agarose. Then the beads were restricted by AluI at 37℃ for 1 h (Figure 2.2.1 step 1), where 1 U AluI were used in transposase method for less frequent cutting compared to 10 U in AluI-A-tailing method. Note that as a restriction enzyme, AluI targets a specific palindrome recognition site (5'-AGCT-3'), and leaves blunt ends after restriction. Restricted DNA fragments were purified from the reaction, while fragments without a biotinylated junction would dissociate from the beads thus were removed during bead washing (Figure 2.2.1 step 2).

The main difference between the two methods was the way the adaptors were ligated to DNA fragments. In AluI-A-tailing method, a free adenine (A) nucleotide was tagged at each 3' end of the fragments (Figure 2.2.1 left, step 3 top), which was used to ligate adaptors with a thymine (T) overhang at their own 3' ends (Figure 2.2.1 left, step 3 bottom). Each adaptor contains a 3-letter barcode, and different barcodes were used to index different libraries (Figure 2.2.1 left, step 3 bottom). During tagmentation, adaptors within the transposase enzymes were inserted at random sites of the DNA fragment (Figure 2.2.1 right, step 3 top). In contrast to the AluI-A-tailing method which takes at least 1.5 h to ligate adaptors, tagmentation only takes five minutes (min) for reaction and another five min for termination. In the tagmented fragment, two different adaptors were tagged to either side of the Hi-C biotinylated junction (Figure 2.2.1 right, step 3 bottom). All libraries used the same adaptors, unlike the AluI-A-tailing method, indexing was carried out in the later PCR step.

In AluI-A-tailing method, before the PCR library amplification, it is critical to remove unligated adaptors as much as possible by repeated washing; otherwise they would be extensively amplified in PCR and affect library quality. Both methods require beads to be transferred to PCR tubes. When adding the primers to the PCR tubes, AluI-A-tailing method used a universal primer mix for all libraries. However, different combinations of PCR primers were used by transposase method to index different libraries, where a maximum of 96 combinations could be achieved by 12 different primer 1 (index i7, Nextera index kit, Illumina) and 8 primer 2 (index i5, Nextera index kit, Illumina) (Figure 2.2.1 right, step 4). AluI-A-tailing method used Platinum Pfx DNA Polymerase for PCR whereas transposase method used KAPA

HiFi polymerase (KapaBiosystems, Figure 2.2.1 step 4). Ideally only the biotinylated fragments are carried by the beads and amplified. The amplified fragments are not biotinylated and thus are bead-free (Figure 2.2.1 step 5). This allowed purification of the amplified library for sequencing. It should be noted that the sequencing for transposase-processed libraries has a specific indexing procedure due to the unique way of indexing libraries, thus could not be pooled with AluI-A-tailing libraries or be sequenced together.

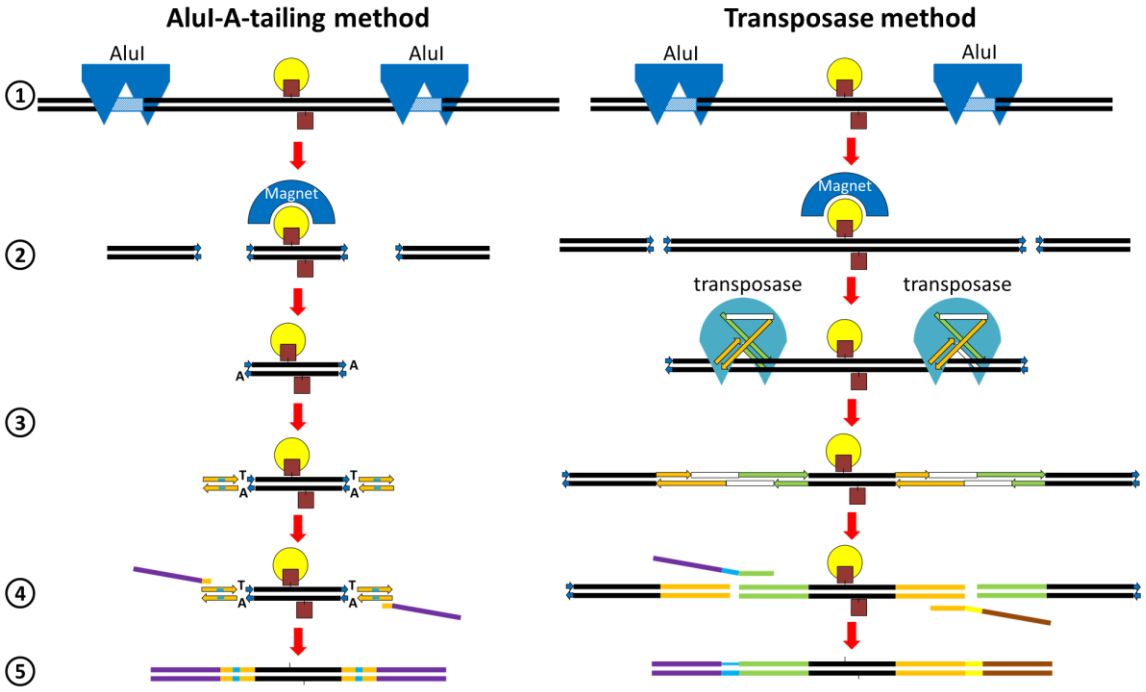


Figure 2.2.1 Schematic of the AluI-A-tailing method (left) and transposase method (right) for preparing sequencing libraries

Steps are numbered according to processing order. Hi-C DNA fragment is shown as black double lines. The yellow circles represent streptavidin-coated magnetic beads which bind to biotin (brown boxes). The blue shaded boxes represent AluI recognition sites on the DNA fragment, which become blunt ended after digestion (short blue arrows). For AluI-A-tailing method (left), pairs of orange arrows represent adaptors ligated via the A-tails, with the 3-letter barcode shown in small pale blue boxes. PCR primers are shown in purple with annealing sequences in yellow. For transposase method (right), two different adaptors on either side of the Hi-C junction on tagmented fragment are shown as green and orange pairs

of arrows. Nextera PCR primer 1 (index i7) annealing to the green adaptor is shown in purple, and primer 2 (index i5) annealing to the orange adaptor is shown in brown, except the complementary sequence in the corresponding adaptor colours. The various barcodes in primer 1 is shown in pale blue and barcodes in primer 2 is in yellow. For both methods the amplified fragment is not biotinylated, but it should still contain the sequences of the Hi-C junction labelled as short vertical lines.

3. Method development – Combined imaging and single-cell Hi-C library preparation using Tn5 transposase

3.1. Chapter introduction

Adding adaptors to Hi-C DNA fragments for PCR primer annealing and library amplification is a crucial step in single-cell Hi-C sequencing library preparation. In the first single cell Hi-C protocol by Nagano et al.[56] this was carried out in three enzymatic steps: (i) cutting DNA into blunt-ended fragments with a restriction enzyme, (ii) attaching an adenine (A) nucleotide to both 3' ends of the fragments using exo-nuclease deficient Klenow, and (iii) attaching adaptors via a thymine (T) overhang on the 3' end of the adaptor using DNA ligase. Each step is not 100% efficient and the inefficiencies accumulate. Also all 3 steps take a minimum of 6 hours and to improve the efficiency as much as possible, overnight (16-20hrs) steps are required. In contrast, the Tn5 transposase system (in the following context, transposase for short) has a number of potential benefits in this adaptor adding process. First, unlike conventional restriction enzymes, transposases do not require a specific recognition or restriction site to fragment DNA molecules, thus reducing sequence bias in the reaction. Second, transposases are capable of catalysing fragmentation and tagging with adaptors simultaneously and efficiently. This combined process, called "tagmentation", can be carried out with current transposase systems in as little as 5 minutes, significantly reducing the time required for library preparation[90]. A detailed comparison of the workflows required for single-cell Hi-C experiments is shown in Section 2.2. Thirdly, especially for studies with limited starting material, a particularly promising factor is that only small amounts of input DNA are required by transposases – as little as 1 ng DNA is suggested by the latest commercial kit, the Nextera XT DNA preparation kit from Illumina[90]. The term "Nextera" and "Nextera XT" in the following context all refer to this kit, although this

is still 3 orders of magnitude more DNA than that found in a single genome (~2.5 pg in a haploid mouse cell).

My first aim was to integrate the transposase system into single-cell Hi-C experiments. In particular, this protocol was developed to carry out single-cell imaging and Hi-C on the same cell (see Section 2.1 for detailed experimental design and workflow). To achieve this, each step of the transposase-based sequencing library preparation needed to be tested with the single cell Hi-C protocol. This chapter reports these tests and all the modifications that have been made to try to improve the tagmentation of single cell Hi-C samples. Modifications that improved the results were implemented right after the corresponding test and included with the following tests or experiments. Also tests in different experimental steps were sometimes carried out in parallel, and the modifications implemented did not follow the order of the steps in the protocol. Results in this chapter are not described in the order of their implementation, but in a more logical order where relevant results are discussed together.

Each full experiment from cell culture to sequencing library quality analysis took two to three weeks, and two more weeks if sequencing is required for DNA sequence analysis. In addition, the success rate of preparing single cells was low especially in the early experiments and many of the reagents used throughout the experiments were expensive. So unfortunately it was usually not feasible to repeat an experiment for a specific modification. But within each experiment, each condition was tested using at least one duplicate sample with controls.

I used high sensitivity DNA chip and Bioanalyzer (Agilent technologies, see method Section 7) to analyse library qualities before sequencing. The main data used were the electropherograms (Figure 3.1.1) and the region tables from DNA smear assays (Table 3.1.1). The electropherogram shows the size distribution pattern of the library in general and roughly indicates the yield by detected fluorescent units (FU). The actual yields within specific size regions were calculated using the concentration data from the corresponding region table and the sample volume. The region table also includes the mean size of DNA within each region, calculated from the smear assay. It should note that the peaks are increasingly inaccurate as size increases due to the

electrophoresis mechanism. Also my libraries were very unlikely to contain DNA under 50 bp, where the lower marker of the chip was set to be 35 bp. So I only selected regions between 50 and 3000 bp for analysis.

Only the fragments within a 300 – 700 base pair (bp) range are size-selected and sequenced. This 300 bp lower limit is used mainly because a Hi-C contact pair is only valid when both end sequences can be uniquely mapped to different regions in the genome. Fragments shorter than 300 bp are likely to have a Hi-C junction too close to one end, such that the short sequence may be unmappable. On the other hand the higher limit is set by the sequencing method (Illumina SBS in our case), where progressively longer fragments do not efficiently form clusters on the flow cell and thus do not provide good sequence data. A good single-cell Hi-C library should have most of its fragments falling into this critical range. This is compatible with the Nextera XT kit as normally it produces DNA fragments in the size range from 150 to 2000 bp[90] as is shown in Figure 3.1.1 c. The size distribution patterns of two good single-cell Hi-C libraries, in which sufficient Hi-C junctions were identified using the transposase method, were used as a sign of good library quality for analysis in other experiments (Figure 3.1.1 a, b). These libraries both have over 60% fragments in sizes between 300 and 700 bp (Table 3.1.1).

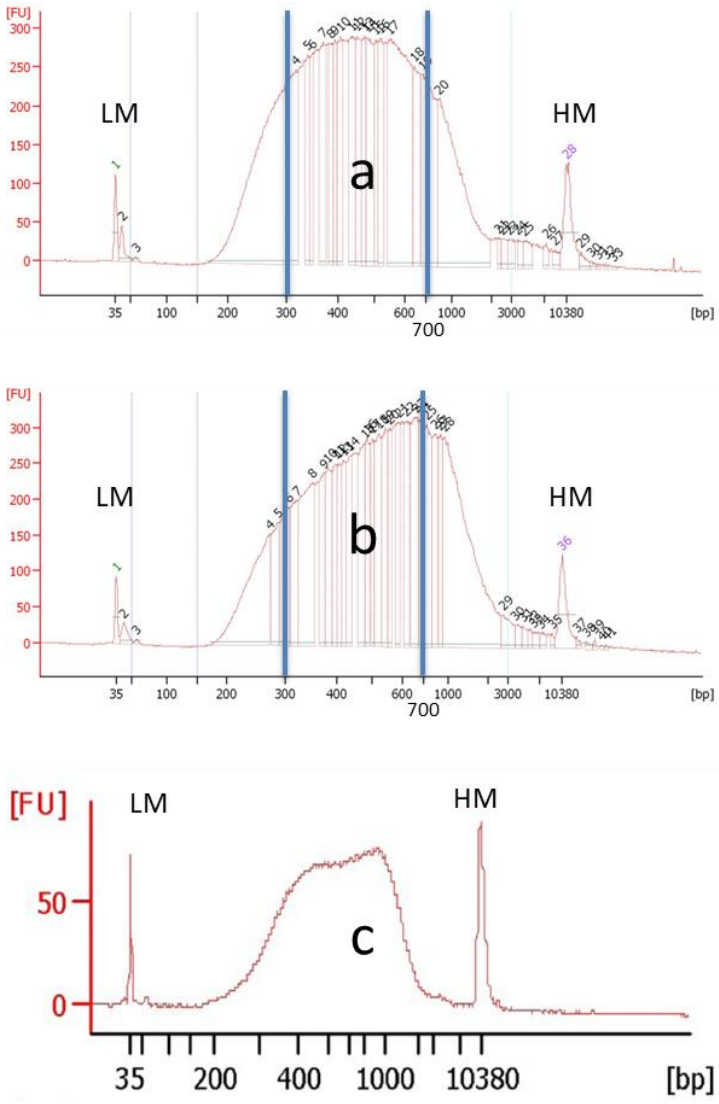


Figure 3.1.1 Size distribution patterns of good libraries

Electropherograms of amplified single-cell Hi-C libraries with the best quality (a, b) and a successful library exemplified in the Nextera XT manual[90] (c) were analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer. The area between the blue lines indicates the range of fragments valid for both sequencing and genome mapping (from 300 bp to 700 bp). LM and HM indicate the lower and higher molecular weight markers of the chip respectively. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. (Single nuclei were processed and then tagmented on-bead by 1/100 relative concentration of transposase as compared to that used in the Nextera XT kit.)

Table 3.1.1 Size distribution parameters of good libraries

a, b) Region tables of amplified single-cell Hi-C libraries shown in Figure 3.1.1 a and b respectively, from the smear assays of high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer. The first two columns indicate the size limits of the manually selected regions. Corr. Area indicates the area under the peak curve within the corresponding region. % of total indicates the percentage of total area that is defined within the region. Average size is the mean size of DNA within the region. Conc. and molarity are the mass and molar concentrations of DNA within the region respectively. Color indicates the colour of vertical separator lines of the region shown in the corresponding electropherograms (see Figure 3.1.1 a and b).

a

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Color
50	3,000	11,095.2	97	549	6,759.96	25,777.5	
50	150	12.9	0	56	11.29	304.1	
150	300	2,017.5	18	261	1,428.28	8,346.1	
300	700	7,522.7	66	465	4,534.46	15,945.7	
700	3,000	1,542.1	13	1,154	785.92	1,181.5	

b

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Color
50	3,000	11,207.3	97	642	7,269.76	25,365.9	
50	150	12.7	0	56	12.17	331.2	
150	300	1,596.0	14	260	1,235.54	7,219.6	
300	700	7,103.4	61	481	4,646.58	15,835.4	
700	3,000	2,495.1	22	1,196	1,375.47	1,979.6	

3.2. Testing the effect of embedding cells in agarose on the tagmentation activity of the transposase

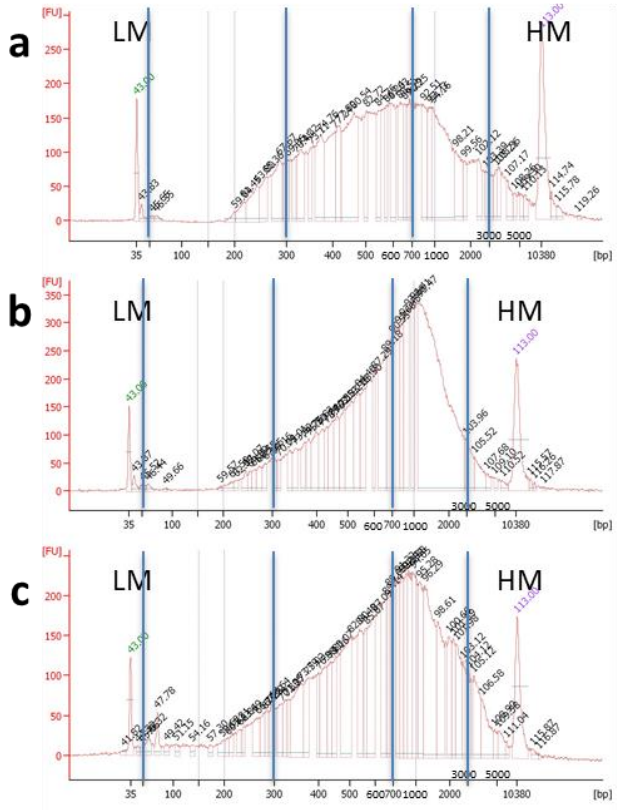
In order to image the single cells on a multi-well plate prior to biochemical single-cell Hi-C processing, my colleagues and I found that it was crucial to cover each cell with an agarose pad made from low melting point (LMP) agarose. This was performed by adding 10 µL 1% LMP agarose solution to each cell in the 384-well plate after single-cell isolation, then leaving at room temperature (RT) until the agarose congealed. This pad trapped and immobilized the cell at the bottom of its well[66,92], largely improved the possibility (roughly from less than 50% to over 80% in average)

and speed (roughly from more than 5 min per cell to less than 2 min in average) of finding a cell in a well during microscopy imaging. The less time required overall also helped prevent the cells from drying out and their DNA structures from denaturing. It also allowed subsequent Hi-C processing by enabling reaction solution exchange without disturbing or losing the cell. This strategy worked because the low melting agarose pad stayed congealed during the 37˚C reaction steps (see Section 2.1). It wasn't until the final crosslink removal at 65℃ that the agarose pad was melted. However, it was still not feasible to remove agarose from the solution until the biotinylated Hi-C DNA had been bound to streptavidin-coated magnetic beads (see Section 2.1). If tagmentation of biotinylated Hi-C DNA was to be carried out before streptavidin purification, the effects of agarose on the reaction need to be tested.

3.2.1. Agarose in the tagmentation reaction of population Hi-C DNA resulted in longer DNA fragments.

I tested the effects of agarose on the tagmentation reaction using 2.5 pg of purified population Hi-C DNA. The samples were dissolved in different concentrations of liquid-state LMP agarose before tagmentation. The final concentration of LMP agarose used in the single cell Hi-C protocol is between 0.5 and 1 percent (0.5% before the crosslink removal step at 65℃ overnight, concentrated up to 1% due to solvent evaporation, actual concentration varied and volume was too low to be feasibly quantified), so I decided to test the effect of no agarose and agarose at 0.5 and 1 percent. All the samples were processed using standard Nextera XT kit protocol, except that the input DNA solution volume was 20 µl instead of 5 µL (used in the kit) to simulate the volume of the single cell samples after cross link removal. The resulting LMP agarose solution was kept at temperatures over 37℃ until the end of streptavidin bead-binding, and was then removed after bead separation. Finally, the purified biotinylated Hi-C DNA was amplified by PCR and the subsequent fragments analysed by high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer (Figure

3.2.1.1). DNA fragments from the no agarose control sample (Figure 3.2.1.1 a) have sizes concentrated between 300 and 700 bp (56% of total), as is expected for a successful tagmentation. However, the sample with 0.5% agarose have less fragments in the 300 – 700 bp range (45%), and more fragments over 700 bp (34% compared with 26% without agarose) (Figure 3.2.1.1 b). Increasing the agarose concentration to 1% resulted in an even greater proportion of fragments over 700 bp (44%) (Figure 3.2.1.1 c). The increase in overall fragment size with higher agarose concentration is also reflected by the increase in mean fragment size between 50 and 3000 bp (742, 854 and 920 bp in 0%, 0.5% and 1% agarose samples respectively) (Figure 3.2.1.1). Therefore, if agarose is present in the transposase reaction, the transposase cuts and incorporates adaptors less frequently, resulting in libraries with longer DNA fragments. These results suggest that agarose inhibits transposase activity, and the more concentrated the agarose the stronger the inhibition. The standard agarose concentration in the protocol is estimated to be between 0.5% and 1% (0.5% before overnight crosslink removal at 65℃ and unmeasurable afterwards due to solvent evaporation), which resulted in a library with increased overall fragment size and less than 50% fragments in the 300 – 700 bp size range. As the fragments between 300 and 700 bp are normally processed for sequencing and Hi-C contact identification, the 0.5% – 1% agarose present in tagmentation was not deemed acceptable.



	Agarose conc.	% of Total 300 – 700 bp	% of Total 700 – 3000 bp	Average size (bp) 50 – 3000 bp
a	0%	56	26	742
b	0.5%	45	34	854
c	1%	45	44	920

Figure 3.2.1.1 An increase in longer DNA fragments is obtained with higher agarose concentrations present in tagmentation.

Electropherograms of tagmented libraries made with no agarose (a), 0.5% agarose (b) and 1% agarose (c) as analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer. The values on the right LM and HM indicate the lower and higher molecular weight markers of the chip respectively. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. % of total indicate the areas under the peak curve within regions 300 – 700 bp and 700 – 3000 bp of the corresponding electropherogram. Average size represents the mean fragment size of the libraries calculated between 50 and 3000 bp of the corresponding library. (All samples contained 2.5 pg of population Hi-C DNA, tagmented with 5 µl of Nextera transposase. Samples were then purified with streptavidin-coated magnetic beads and amplified with 20 cycles of PCR using Nextera PCR mix.)

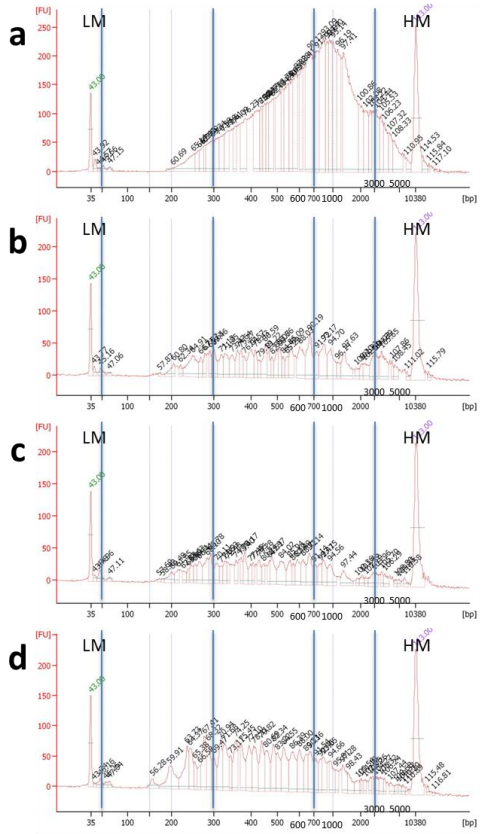
3.2.2. Single cell Hi-C nuclei processed with agarose could not be properly tagmented.

To further investigate the effects of agarose on tagmentation, I repeated the same test outlined above in Section 3.2.1, but this time I assayed single-cell Hi-C nuclei along with 2.5 pg control population Hi-C DNA. Again, for each sample, the 20 µL input volume was greater than the suggested 5 µL, and streptavidin-coated bead purification of the DNA was carried out after tagmentation. After PCR amplification the library fragments were analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer (Figure 3.2.2.1). The increased proportion of larger DNA fragments of the control library largely resembled the library processed under the same condition in the previous test (compare Figure 3.2.1.1 c with Figure 3.2.2.1 a). This confined good sample and reagent conditions in this experiment and allowed a direct comparison with the previous test. Interestingly, for the single-cell libraries, the DNA fragment distribution was not the same as the population Hi-C library (Figure 3.2.2.1 compare a with b - d). Instead of a high proportion of large fragments the single cell Hi-C libraries contained fragments enriched in the optimum range of 300 to 700 bp. However, the patterns of all the single cell plots contained many obvious spikes instead of a smooth line as observed with the control population Hi-C sample. The yields of the single cell libraries were also 2 to 3 times lower than the control sample.

To further investigate the success of the transposase reactions I decided to send the three single cell nuclei libraries along with the control population Hi-C library for sequencing. Analysis of the sequencing data using NucProcess, a python program designed to identify valid Hi-C contacts showed very low sequence variety (Table 3.2.2.1). This indicates that only a few fragments were amplified to form the library, which means that tagmentation in these single-cell samples only occurred on limited occasions. Also the sequences were distributed over the whole of the genome contact map (Figure 3.2.2.2), which suggests the low coverage was not due to defects in the

input DNA. So in addition to the inhibition at tagmentation, the agarose pad caused Hi-C reaction deficiencies for single nucleus samples. One possible explanation could be that agarose made the DNA more inaccessible to the reagents. This concept of accessibility seemed to be the common rationale of several implemented modifications on the transposase method and will be further discussed in Section 6.1.1.

In general, agarose is not compatible with tagmentation and should be removed before the reaction. More effects and consequences of agarose in input DNA will be further discussed in Section 6.1.2.



	Sample type	Agarose conc.	% of Total 300 – 700 bp	% of Total 700 – 3000 bp	Average size (bp) 50 – 3000 bp	Yield 300 – 700 bp
a	2.5 pg DNA	0.5%	49	36	863	41.9 ng
b	single nucleus	0.5%	48	20	679	13.0 ng
c	single nucleus	0.5%	50	15	591	12.1 ng
d	single nucleus	0.5%	52	15	584	22.4 ng

Figure 3.2.2.1 Agarose inhibits tagmentation of single-cell Hi-C nuclei

Electropherograms of tagmented libraries made with 2.5 pg population Hi-C DNA (a) and three single cell Hi-C nuclei (b, c, d) as analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer. LM and HM indicate the lower and higher molecular weight

markers of the chip respectively. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. % of total indicate the areas under the peak curve within regions 300 – 700 bp and 700 – 3000 bp of the corresponding electropherogram. Average size represents the mean fragment size of the libraries calculated between 50 and 3000 bp of the corresponding library. Yield in 300 – 700 bp range is calculated using the concentration and purified library sample volume (All input samples were in 20 µL of 0.5% low-melting-point agarose. All samples were tagmented with 5 µl of Nextera transposase. Then all samples were purified by streptavidin-coated magnetic beads. All libraries were amplified using 20 cycles of PCR using Nextera PCR mix.)

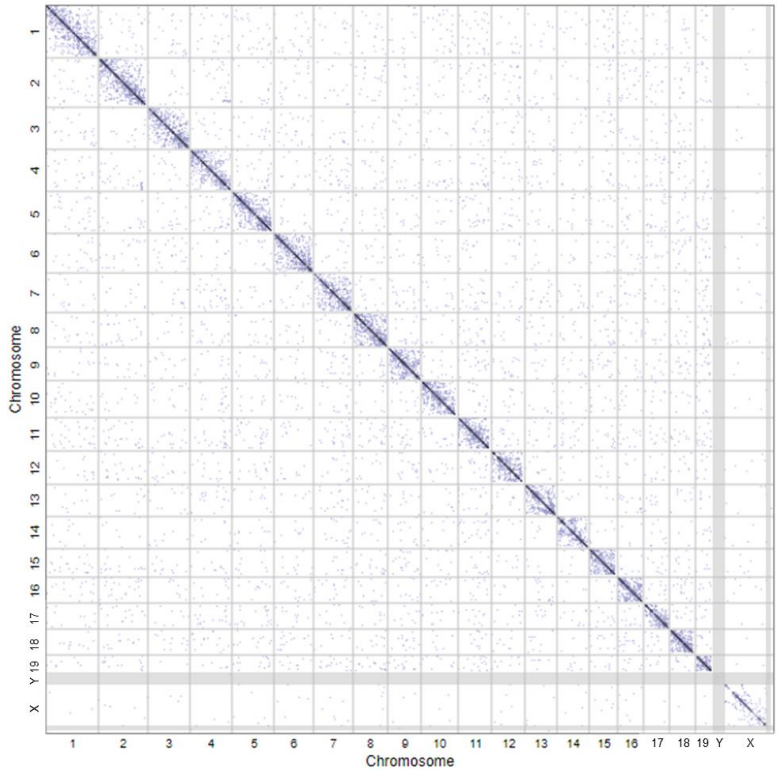


Figure 3.2.2.2 Sequence analysis of library tagmented with agarose present

Hi-C contact map of the control sample in Figure 3.2.2.1 a.

Table 3.2.2.1 Sequencing read analysis on libraries shown in Figure 3.2.2.1

Cells in Fig. 3.2.2.1	Input read pairs[a]	Uniquely mapped pairs[b]	Filtering accepted[c]	Deduplicated total contacts[d]	Cis contacts[e]	Trans contacts[f]
a[g]	655,979	320,891	148,262	27,950	25,451	2,499
b	783	408	58	8	8	0
c	2199	388	31	2	2	0
d	10,334	6,136	620	105	101	4

[a] Total number of paired-end reads for each sample.

[b] Hi-C contact read pairs that map to unique positions in the reference genome.

[c] The number of Hi-C contacts after filtering contacts uninformative for structure.

[d] Total contacts after removing PCR sequencing duplicates.

[e] Intrachromosomal contacts.

[f] Interchromosomal contacts.

[g] Control sample with 2.5 pg population Hi-C DNA, whose contact profiles resemble a good single-cell Hi-C library.

3.3. Investigating the transposase reaction with Hi-C DNA bound to beads

When preparing single cell Hi-C libraries it is critical to purify biotin-labelled DNA fragments away from non-biotinylated DNA as this enriches the library with sequences containing valid Hi-C contacts. This is done by binding the biotin-labelled DNA fragments to streptavidin-coated magnetic beads (Figure 2.1.1). Thus, virtually all reactions during the library prep stage are carried out on beads. As long as the DNA fragments are captured on the beads, they are impossible to elute until the PCR amplification step when the DNA is denatured and bead-free amplicons are formed. So the reactions that need to be carried out off the magnetic beads are those before the bead binding step (see Figure 2.1.1). These involve the removal of the formaldehyde

cross links in the DNA at 65℃ (see Figure 2.1.1), which melts and dissolves the agarose and Hi-C DNA in PBS buffer. However, this solution also poses two possible problems. Firstly, the single-genome amount of DNA is most likely further reduced due to the purification step, to an unknown amount. This is particularly critical for tagmentation as the amount of transposase enzyme relative to the amount of input DNA is an important factor for generating libraries of the correct fragment size (as will be discussed in Section 3.3.1 – 3.3.3). Secondly, and maybe more importantly, we did not know what effect the DNA bound to streptavidin on magnetic beads will have on the tagmentation reaction, and this needed to be carefully tested.

3.3.1. An excess of transposase enzyme over-cuts Hi-C DNA bound to magnetic beads.

In the initial experiments single-cell Hi-C libraries were processed using the exact procedure described in the Nextera XT kit manual, where the amount of transposase needed to process 1 ng of DNA was used with Hi-C DNA from a single nucleus bound to beads. The PCR-amplified library from these reactions were analysed on high-sensitivity DNA chips with Agilent 2100 Bioanalyzer and a representative trace is shown in Figure 3.3.1.1 a. The size distribution of DNA fragments from these libraries were unusual as they displayed a sharp peak at ~400 bp and two “shoulders” at ~200 bp and ~1000 bp respectively.

As I outlined earlier the optimum DNA fragment size required for high-throughput DNA sequencing is within the range of 300 to 700 bp. Hence both the smaller and larger fragments containing the 200 bp and 1000 bp shoulders were mostly removed by size selection before sequencing (Figure 3.3.1.1 a).

After analysing the sequencing data, most identified amplicons in the on-bead tagmented libraries were shown to contain short pieces of the Nextera adaptor DNA and primer sequences rather than the expected mouse DNA fragment pairs of the Hi-C DNA (Figure 3.3.1.2). These sequences were obviously not suitable for mapping to

the mouse genome or acting as Hi-C contacts. This suggested that, because the single cell input DNA was at a far lower amount than the suggested 1 ng in the Nextera kit, after the first tagmentation the successfully inserted Nextera adaptors were most likely being repetitively tagmented by the excess transposase, resulting in approximately 400 bp fragments (Figure 3.3.1.1 b). This distance could be due to the physical shape of the streptavidin-coated magnetic beads preventing the transposase from targeting regions closer to the Hi-C junction. This facilitates repeat tagmentation at the observed distance. The inaccessibility issues will be further discussed in Section 6.1.1.

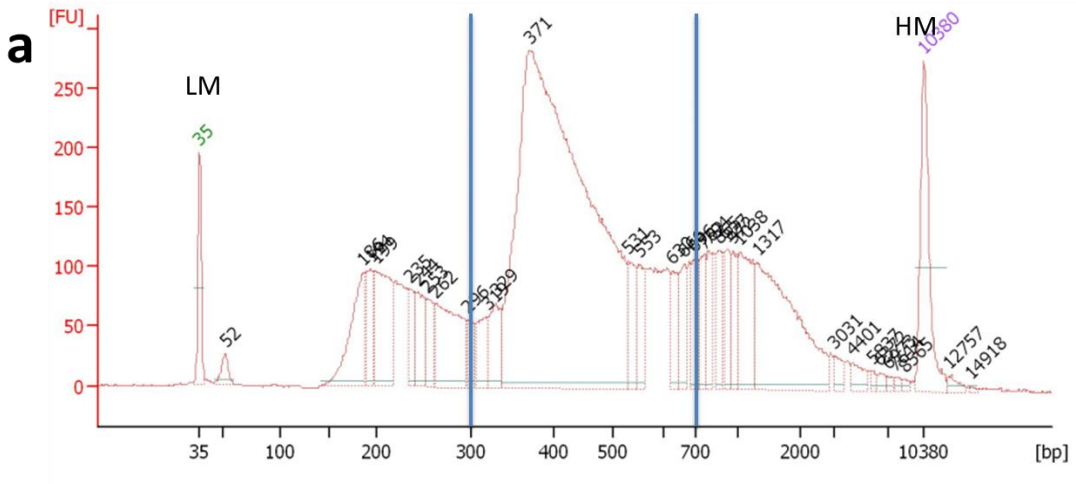


Figure 3.3.1.1 Over-tagmentation of single cell biotinylated Hi-C DNA bound to streptavidin beads

(See the next page for the rest of the figure and figure legends)

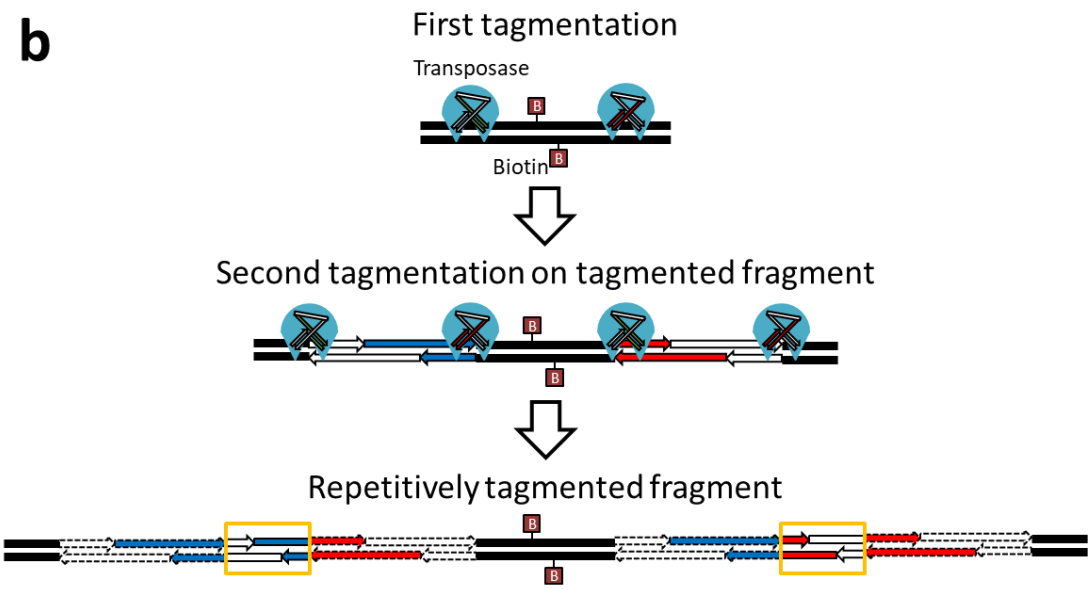


Figure 3.3.1.2 Over-tagmentation of single cell biotinylated Hi-C DNA bound to streptavidin beads

a) Typical size distribution of over-tagmented single-cell Hi-C libraries. The pattern shown is an electropherogram of an amplified library analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. b) Schematic of conjectural over-tagmentation mechanism. Primer 1 and 2 are labelled in green and red respectively. The solid and dotted arrows represent primers used in the first and second tagmentation respectively. The orange box indicates the bits of primer sequence as the resultant reads if the bottom fragment is amplified and sequenced. (Single nuclei were tagmented on-bead at concentration suggested in the Nextera XT kit.) The area between the blue lines indicates the range of fragments valid for both sequencing and genome mapping (from 300 bp to 700 bp).

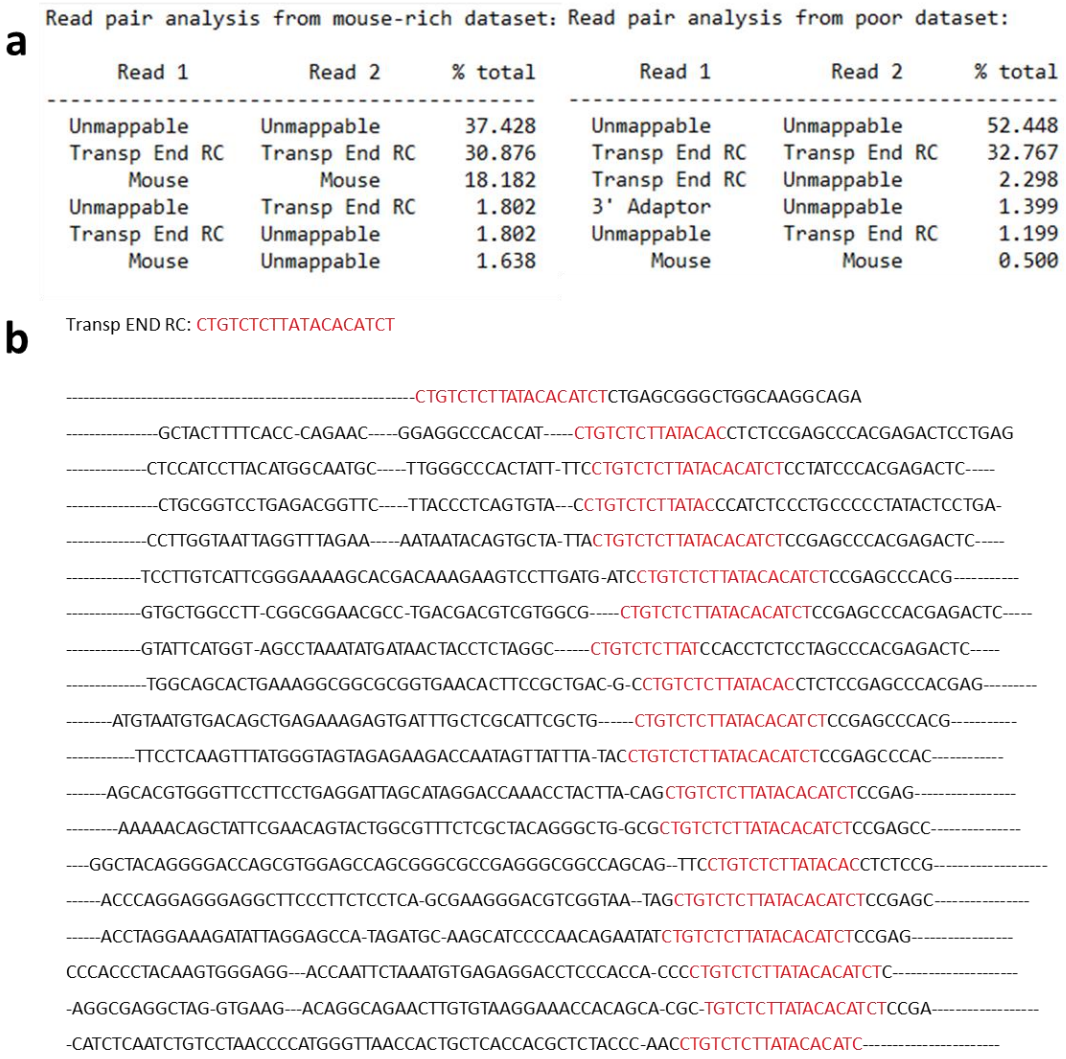


Figure 3.3.1.2 Sequence read analysis showing inserted pieces of Nextera primer sequences

a) A summary of read analysis on randomly selected 1000 reads from a good dataset with relatively more mappable mouse DNA (top) and a poor dataset with relatively less mouse DNA (bottom). Both datasets were from over-tagmented samples. “Transp END RC” is the end part of the Nextera Index Read Primer reverse complementary sequence (CTGTCTCTTATACACATCT). b) Example reads from the mouse-rich dataset, with full or partial hits on the Transp END RC highlighted in red. (Data and analysis were provided by Dr. Tim Stevens.)

My experiments suggested that the original Nextera transposase reaction conditions are not useful for single-cell amounts of Hi-C DNA bound to beads. The unusual size

distribution pattern of the resultant libraries (Figure 3.3.1.1 a) was thought to be a typical sign of over-tagmentation. It thus became evident that I would need to carefully optimise various conditions of the tagmentation reaction for on-bead single-cell amounts of Hi-C DNA.

3.3.2. Trimming the DNA before tagmentation is critical for preparation of single cell libraries.

Sequence analysis of the transposase-processed libraries suggested that the libraries always had a high number of uninformative amplicons such that only one DNA fragment rather than a pair of fragments ligated via a Hi-C junction were identified. This suggests that the tagmentation did not cut the DNA off to form smaller fragments. Instead, the two adjacent ligated DNA fragments were linked by either the unremoved transposase complex or a shared adapter sequence, or both. Unfortunately this mechanism cannot be confirmed because no molecular details about Nextera transposase reaction are available, but it is consistent with the DNA transposition mechanism of the natural Tn5 transposase[84,93]. Consequently, the fragments without a Hi-C junction could not be purified by means of biotin and streptavidin beads during the library preparation process, and would be amplified along with valid Hi-C fragments to constitute the library (Figure 3.3.2.1 a top). The sequences of these amplicons contain no information about genome structure, so they reduce library quality and occupy a significant portion of sequencing capacity. This mechanism will be further discussed in Section 6.1.

An alternative strategy is to use the Nextera transposase simply to add the adaptors, which requires that the DNA should be fragmented first. Then, after the biotin purification, only shorter fragments containing biotinylated Hi-C junction DNA proceed to tagmentation. In this scenario it would be much less likely to have as many uninformative fragments in the library (Figure 3.3.2.1 a bottom). In addition, it is probably more important to use this strategy to open the rather closed conformation of

the single genome DNA. Based on most of my experiments using the transposase method, tagmentation on single-genome-equivalent amounts of DNA (2.5 pg) from population Hi-C control samples was always more consistent than DNA samples from single nuclei (data not shown). It was thought to be the dissolving level that may lead to the difference in library quality between the two types of samples. The population Hi-C DNA is already fragmented to some degree when precipitated and purified away from proteins and RNAs, and dissolved in solution buffer; whereas DNA from single nuclei will most likely still be bound to some proteins and RNAs and be less fragmented. So the DNA from single nuclei will more likely be in a less accessible structure when bound to the streptavidin-coated magnetic beads. This less accessible structure would not only possibly reduce the efficiency of bead purification, but also make the interior DNA inaccessible to transposase. Therefore, I decided to investigate if trimming the DNA using the restriction enzyme AluI, would make the DNA more open and accessible to transposase enzyme.

Interestingly, when single nuclei Hi-C DNA bound to magnetic beads was tagmented with a reduced amount of transposase (this experiment was done after the transposase amount adjustment experiment, see Section 3.3.3 for more discussion), comparison between libraries made from trimmed and untrimmed Hi-C DNA showed that the trimming removed the unusual size distribution pattern due to over-tagmentation (Figure 3.3.2.1 b). Instead trimming the Hi-C DNA resulted in a more even distribution of fragments between 300 and 700bp. I found that this approach gave more consistent results regardless of the sample quality. Although the exact mechanism behind this improvement needs further investigation, it was thought that trimming opens the DNA conformation, and provides more consistently sized DNA samples. As it was mentioned in Section 3.2.2, this concept of accessibility and uniform size of DNA samples seem to be very important in the transposase-based single-cell Hi-C library processing method. This will be further discussed in Section 6.1.1.

In the only experiment that successfully produced two good libraries, the DNA conformation was most likely in a more open conformation. So in that experiment

most of the samples were efficiently tagmented. The reason behind this is still unknown, but single-nucleus samples in the experiment were processed either trimmed or untrimmed before tagmentation. Similar to the samples shown in Figure 3.3.2.1 b, these samples were also processed using a reduced amount of transposase to avoid over-tagmentation (see Section 3.3.3 for more discussion). As shown in Figure 3.3.2.1 c, trimmed libraries had smaller fragment size distributions. This is probably because the biotin purification after trimming further reduces the amount of DNA in a single nucleus sample, by removing an unknown amount of the fragments without a biotinylated Hi-C junction. It would also be possible to shift the size distribution of untrimmed libraries to the small side by adding more transposase, but this would increase the risk of over-tagmentation.

In my optimisation, trimming was carried out using the AluI restriction enzyme, which acts as the second restriction step in the original AluI-A-tailing single-cell Hi-C method. However a much lower amount of AluI (1 unit compared to 10 units used in AluI-A-tailing method) was used during the same 1 hour incubation at 37℃. This trimming method worked well to give libraries with good quality.

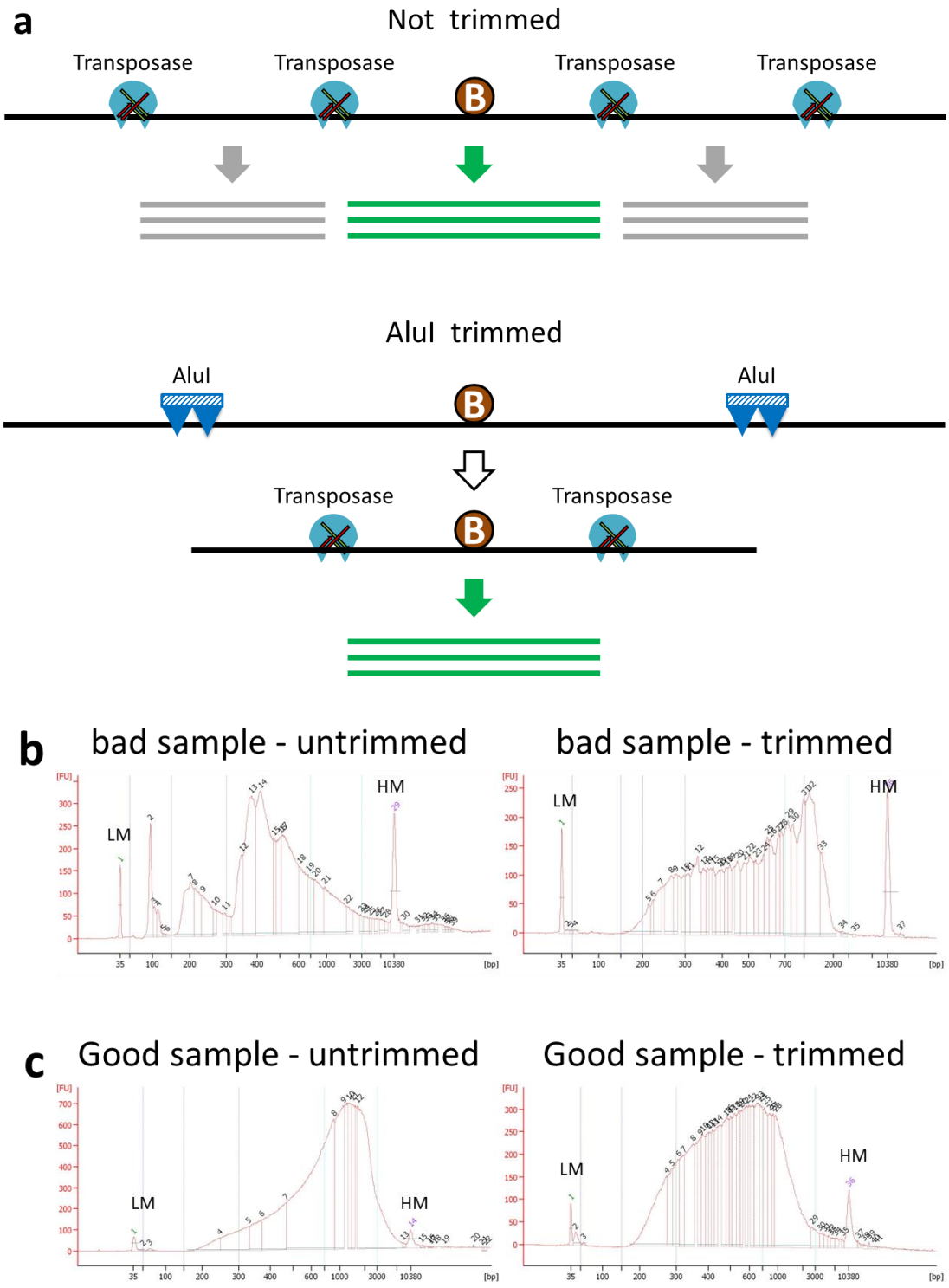


Figure 3.3.2.1 Comparison between untrimmed and AluI-trimmed workflows for transposase-based single-cell Hi-C library preparation

a) Schematics of the two workflows for non-trimmed and AluI trimmed Hi-C DNA. Biotin labels on Hi-C junctions are shown as brown circles. In amplified libraries, fragments with Hi-C junction are shown in green whereas fragments without a Hi-C junction are shown in

grey. b & c) Electropherograms of amplified single-cell libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer, showing size distribution patterns and relative yields of libraries processed with and without AluI trimming. The two libraries in b are examples from failed experiments. The two libraries in c are examples from the only successful experiment where sufficient Hi-C junctions were identified. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower maker and higher molecular weight markers of the chip respectively. (All single nuclei were tagmented on-bead using 1/100 dilution of Nextera transposase. All libraries were amplified with KAPA HiFi DNA polymerase with 25 cycles of PCR (modifications in the PCR reaction will be discussed in Section 3.4).)

3.3.3. Determining the optimal transposase concentration and reaction volume for single nuclei amounts of Hi-C DNA bound to beads

As discussed in Section 3.3.2, trimming the DNA can remove the unusual size distribution pattern due to over-tagmentation of the Hi-C DNA bound to beads. However, even with promising size distribution patterns, sequence analysis of some libraries still showed signs of over-tagmentation (Figure 3.3.3.1 compare a with b). To avoid over-tagmentation, reducing the amount of transposase in the tagmentation reaction was the most logical optimisation to consider. As transposase randomly targets DNA, the relative amount of transposase compared with the amount of input DNA largely determines the resultant sequence length. When the Nextera transposase was used, our single nucleus input (~2.5 pg DNA) was about 400 times less than the suggested input amount (1 ng). This indicates that the input transposase is 400 times in excess for our single-cell Hi-C samples, but diluted transposase could have reduced

activity. On each target DNA fragment, tagmentation is required on both sides of the biotinylated Hi-C junction to ligate adapters for amplification. Inefficient tagmentation, where only one or neither side was tagmented, would result in no amplification and loss of Hi-C contacts. This would reduce the complexity of the resultant library and the quality of reconstituted genome structure. So next I set out to find an optimal amount of transposase for single-cell Hi-C experiments, avoiding both over-tagmentation and inefficient tagmentation. It should be noted that the exact concentration of transposase in the Nextera XT kit is not made available by the company (Illumina). So the term "amount" here refers to the relative suggested concentration of the transposase stock (named ATM, amplicon tagment mix, in the kit) for 1 ng of input DNA.

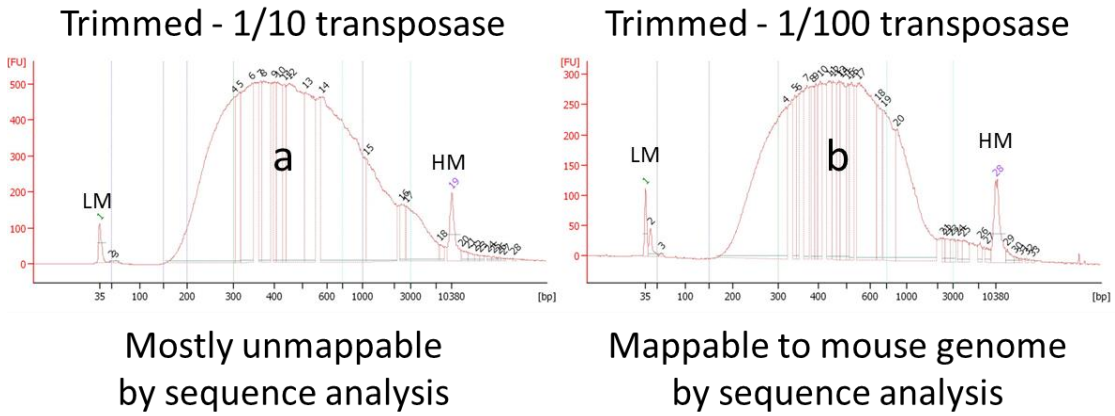


Figure 3.3.3.1 The range of DNA fragment sizes of a trimmed library is not a good indicator of over-tagmentation.

Electropherograms of amplified libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. Library a was tagmented with 1/10 dilution of Nextera transposase (ATM) whereas library b was tagmented with 1/100 dilution. Sequence analysis data not shown. (Both libraries were trimmed single-cell Hi-C samples and were amplified with 25 cycles of PCR by KAPA HiFi

polymerase (modifications in the PCR reaction will be discussed in Section 3.4).)

To determine the optimal amount of transposase to use for single nuclei, three "titration" experiments were carried out using population Hi-C DNA as test samples. The first experiment tested the most suitable concentration of transposase for different amounts of input DNA and was carried out in solution without DNA bound to beads. This was mainly to investigate conditions that would generate the required size distribution pattern of fragments. Such conditions could then act as references and controls in later tests using single-genome-equivalent amounts of DNA. I tested four different amounts of input DNA (1 ng, 100 pg, 20 pg and 2.5 pg single-genome equivalent) with up to four different relative concentrations of transposase (1, 1/10, 1/50 and 1/400 dilution) with the results shown in Figure 3.3.3.2. As expected when 1ng of DNA was processed with the amount of transposase suggested by the manufacturer, an even distribution of fragments between 200 and 1000 bp was obtained (Figure 3.3.3.2 a). The expected distribution of fragments was also found with the concentration of transposese for 100 pg (1/10 dilution) and 20 pg (1/50 dilution) of DNA respectively (Figure 3.3.3.2 c & f). However, this scaling no longer worked for 2.5 pg DNA, which required a 1/50 dilution of transposase rather than the predicted 1/400 to produce a relatively good distribution of DNA fragments (Figure 3.3.3.2 compare i with j). When too high a concentration of transposase was used the resultant libraries for 100, 20 and 2.5 pg of DNA had fragments of a considerably smaller size, which suggested these libraries had been over tagmented (Figure 3.3.3.2 b, d, e, g & h). This test indicates that a 1/400 dilution of transposase is too diluted to achieve efficient tagmentation of single genome amounts of DNA. The 1/50 distribution pattern has a slight increase in smaller 150 bp fragments, indicating some over-tagmentation and a slight excess of transposase, so further tests were needed to investigate the optimal dilution of between 1/50 and 1/400.

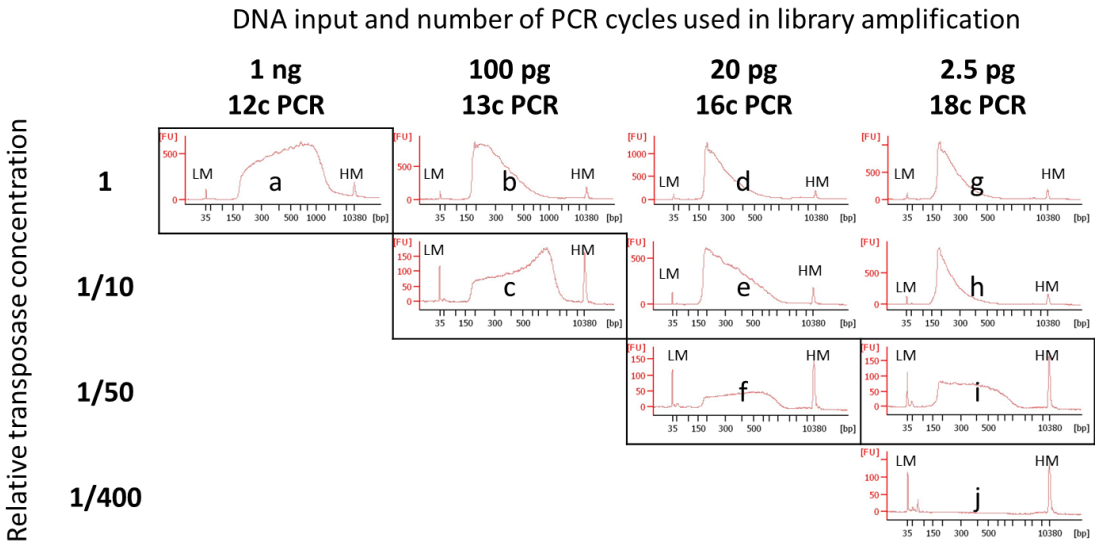


Figure 3.3.3.2 Transposase titration test with varying amounts of input DNA

Electropherograms of amplified libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer, showing size distribution patterns of libraries processed by different conditions. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. (All samples were population Hi-C DNA, tagmented without beads in solution.) 10 libraries a – j are aligned according to their input DNA amounts (horizontally) and relative transposase concentrations (vertically). Reduced input DNA was amplified with more PCR cycles to compensate for the lack of yield in the amplified libraries. The square boxes indicate libraries with the best size distribution pattern for that amount of input DNA.

The second titration test investigates more precise optimal dilutions of transposase for 2.5 pg of DNA. Samples were processed using five different relative concentrations of transposase from 1/25 to 1/400 with the results shown in Figure 3.3.3.3. The distribution pattern of fragments showed a clear decrease in tagmentation level as the concentration of transposases decreased: 1/25 and 1/50 transposase over-tagmented; 1/100 transposase tagmented well; 1/200 and 1/400 transposase tagmented inefficiently (Figure 3.3.3.3 a – e). These results suggested that 1/100 relative

concentration of the transposase was the lower limit to efficiently tagment 2.5 pg of DNA. Hence a 1/100 dilution was set to be the optimal relative transposase concentration for our transposase single-cell Hi-C protocol.

In the third test I decided to repeat the second test except this time the tagmentation reaction volume was halved from 20 µL down to 10 µL. This meant the same exact volumes of Nextera transposase stock and the same amount of DNA were used but in half the reaction volume of buffers (TD, tagment DNA buffer, and NT, neutralize tagment buffer) and input DNA solution (from 5 µL to 2.5 µL). In other words, the concentrations of both transposase and DNA were doubled, keeping the absolute amount of transposase and DNA unchanged. The idea was to achieve a transposase-to-DNA concentration closer to the original value suggested by the manufacturer (1 times' transposase enzyme to 1 ng DNA), and to see whether this could improve the tagmentation. The results of halving the reaction volume also showed an obvious decreasing trend of tagmentation level (Figure 3.3.3.3 f – j). The optimal relative transposase concentration in the half volume tagmentation test was also 1/100 (compare boxed panels c and i in Figure 3.3.3.3), but actually used half the reagent stock compared with the normal volume reaction. In addition, comparison between the two 1/100 libraries shows that the half volume library had more even distribution of fragments in the 300 to 700 bp region whereas the normal volume library had a higher proportion of smaller fragments, indicating over-tagmentation. This might be due to the halved transposase-to-DNA ratio in the half volume reaction, which is closer to the ratio suggested by the manufacturer of the Nextera kit. Alternatively, it is also possible that doubling the DNA concentration, slightly compensated for the dilution of the DNA solution. An obvious benefit of the half volume tagmentation reaction is that it used only half the amount of reagents. This is worth mentioning because the Nextera XT kits are quite expensive and each kit is designed to process only 24 samples, while normally more than 20 samples need to be processed in each single-cell Hi-C experiment. Due to technical limits, such as pipetting and the well dimensions of the 384 well plates, it was not possible to reduce the reaction volume further. However, the half volume reaction did improve the result

and it was implemented in the optimised protocol.

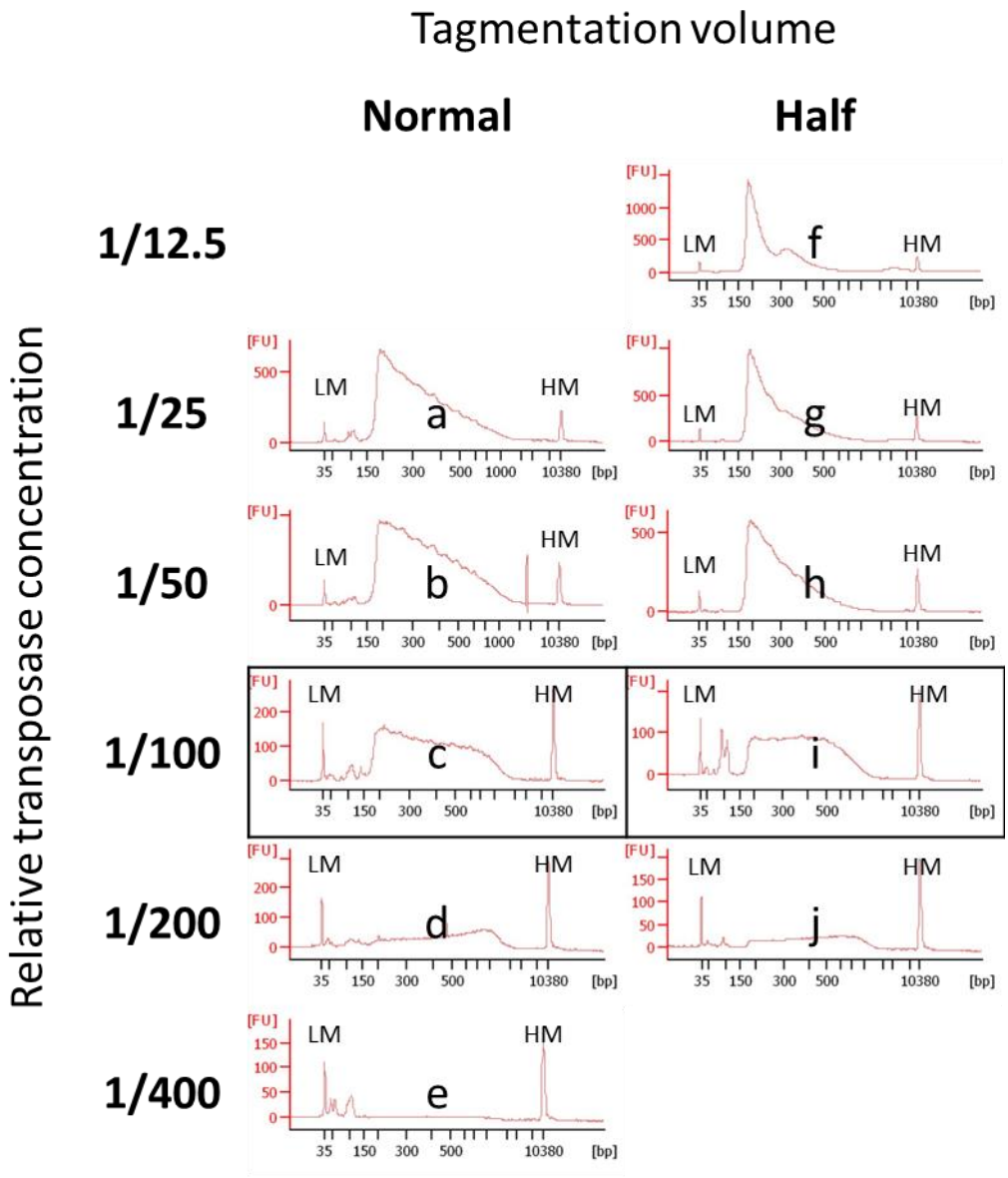


Figure 3.3.3.3 Transposese titration test with 2.5 pg of input DNA and two reaction volumes

Electropherograms of amplified libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer, showing size distribution patterns of libraries tagmented under different conditions. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. (All samples

were population Hi-C DNA, tagmented in solution without binding to beads.) 10 libraries a – j are aligned according to tagmentation reaction volume (horizontally) and relative transposase concentrations (vertically). The square boxes indicate libraries with the best size distribution pattern in the group for a particular reaction volume.

3.4. Optimisations in PCR library amplification

3.4.1. Nextera PCR reagents are not compatible with Hi-C libraries processed on beads.

After optimising the steps for carrying out tagmentation of Hi-C DNA bound to streptavidin-coated magnetic beads, the next question to investigate was the compatibility of different PCR reagents for library amplification. This is important because in the single cell Hi-C protocol the biotin labelled Hi-C DNA undergoes extensive washing and buffer exchange before the PCR step. These specific washing steps are not considered in the Nextera XT manufacturers' protocol where the buffers used in the tagmentation reaction are actually all present in the PCR amplification step. Therefore, it was crucial to test buffer and PCR conditions that were suitable for carrying out the PCR reaction. Another important point to mention is the buffer composition of the reagents included in the Nextera kit, like the tagmentation reaction buffer (TD), termination buffer (NT), transposase enzyme stock (ATM) and Netera PCR master mix (NPM) are not available. Finally, if I was to use the NPM mix from the Nextera kit for PCR, then after washing I had to add all the subsequent buffers back to the sample to maintain the correct buffer environment for NPM reaction. However, I did not want to add active transposase-containing ATM back to the processed single cell, because of the risk of re-tagmentation (see Figure 2.2.2).

I tested the effects of adding different combinations of buffers back to the Nextera NPM PCR reaction for 2.5 pg of Hi-C DNA tagmented and purified on beads and compared it to a control sample tagmented in solution without bead purification

(Figure 3.4.1.1). As previously observed the electropherograms of the amplified libraries of the control sample displayed an even distribution of fragments in the 200 to 1000 bp range (Figure 3.4.1.1 a), suggesting a normal level of tagmentation of the DNA. Interestingly, with either the TD or NT reagents were omitted the purified on-bead samples displayed libraries with significant level of unusually large fragments (Figure 3.4.1.1 b – d). However, in the presence of both of these reagents no significant level of fragments was observed suggesting that this amplification was largely inhibited (Figure 3.4.1.1 e). The results from this test indicate that the beads were affecting the Nextera NPM PCR reaction, causing an unusually large shift in the fragment size distribution of the libraries. The reason why the addition of both the TD and NT mixture inhibited the PCR was unclear.

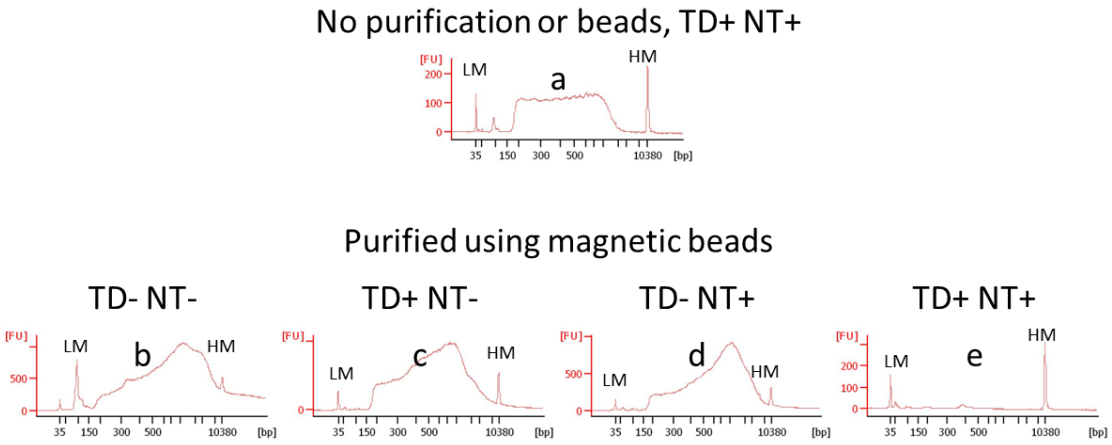


Figure 3.4.1.1 Effects of beads and different buffer conditions on Nextera NPM PCR reaction

Electropherograms of PCR amplified libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer, showing the DNA size distribution patterns of libraries amplified with or without tagmentation reaction buffer (TD) and/or termination buffer (NT). The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. (All samples used 2.5 pg population Hi-C DNA, tagmented by 1/100 ATM in a half-volume reaction without trimming (This test

was done before the implementation of trimming, due to time limit tests with trimming were not carried out.). The number of PCR cycles was 18 for a and 25 for b – e.)

3.4.2. KAPA HiFi DNA polymerase correctly amplified libraries of Hi-C tagmented DNA bound to beads.

It was evident from the previous test that the Nextera NPM PCR system was not suitable for amplifying single cell Hi-C libraries bound to streptavidin beads. Therefore, I investigated other PCR systems to replace the Nextera PCR reagents. Reading the literature revealed that a promising candidate had been previously used in combination with home purified Tn5 transposase that was completely independent from the Nextera kit[89]. The PCR enzyme they used in this procedure was the KAPA HiFi (high fidelity) polymerase, a commercial PCR system with good fidelity characteristics from KAPA Biosystems. I decided to investigate the KAPA HiFi PCR system with the transposase-based single-cell Hi-C method. Using 2.5 pg of population Hi-C DNA tagmented on beads as a sample, I tested whether the tagmentation buffers or Tris buffer alone should be used with the KAPA PCR system. The electropherograms of the amplified libraries showed that the KAPA PCR system only produced an even distribution of DNA fragments in Tris buffer (Figure 3.4.2.1 a). However, amplification of DNA fragments was significantly inhibited by TD buffer and completely inactivated by NT buffer (Figure 3.4.2.1 compare b with c & d), suggesting that neither were compatible with the KAPA PCR system.

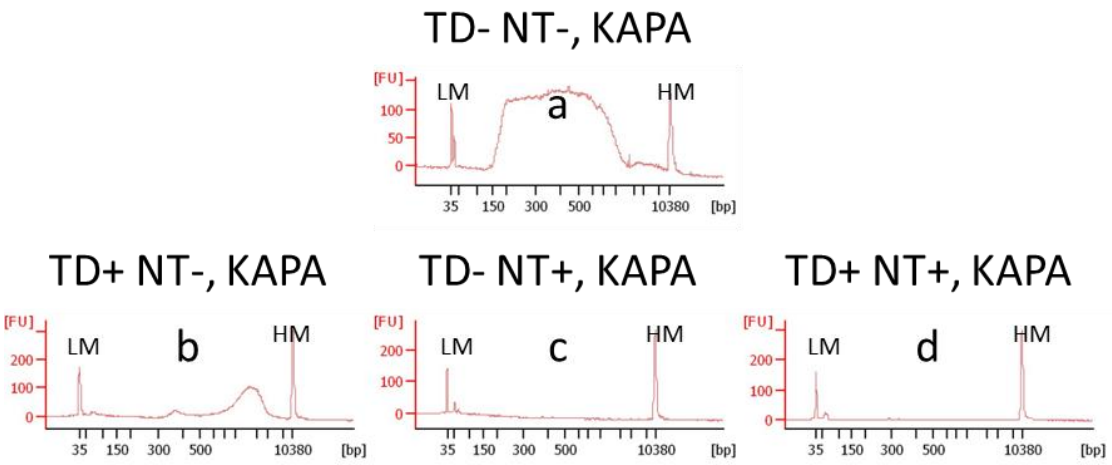


Figure 3.4.2.1 Effects of different tagmentation buffers on KAPA PCR

Electropherograms of PCR amplified libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer, showing the size distribution patterns of libraries amplified with or without tagmentation reaction buffer (TD) and/or termination buffer (NT). The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. (All samples used 2.5 pg population Hi-C DNA and were trimmed using 1 U AluI at 37℃ for 1 hour, tagmented by 1/100 ATM in a half-volume reaction. All libraries were amplified by KAPA HiFi DNA polymerase in 25 cycles of PCR.)

3.4.3. A total of 25 cycles of PCR was required to properly amplify biotin-purified and tagmented DNA from a single genome.

As discussed earlier in Section 3.3.2, single-cell Hi-C DNA was fragmented by restriction enzyme trimming and also when tagmented with the transposase. During the subsequent washes after these steps, the fragments without a biotinylated Hi-C junction should not bind to streptavidin-coated magnetic beads, and thus are washed away from the sample. As a result, the single-genome amount of DNA was further reduced to an unknown amount. To achieve a high enough yield for pooling and

processing up to 24 multiple single cell libraries for Illumina sequencing, we determined that we needed at least 20 ng of 300 – 700 bp fragments per library. In a test using 2.5 pg population Hi-C DNA samples, I found that 25 cycles of PCR, using the KAPA HiFi DNA polymerase, was optimal. This gave a yield that was comparable to that obtained in a control tagmentation of an unpurified library amplified by 18 cycles of PCR using NPM (Figure 3.4.3.1). The successful two libraries, amplified by KAPA and 25 cycles, have good sequence quality overall. This suggests that the increased cycles did not cause significant complexity issues or sequence bias.

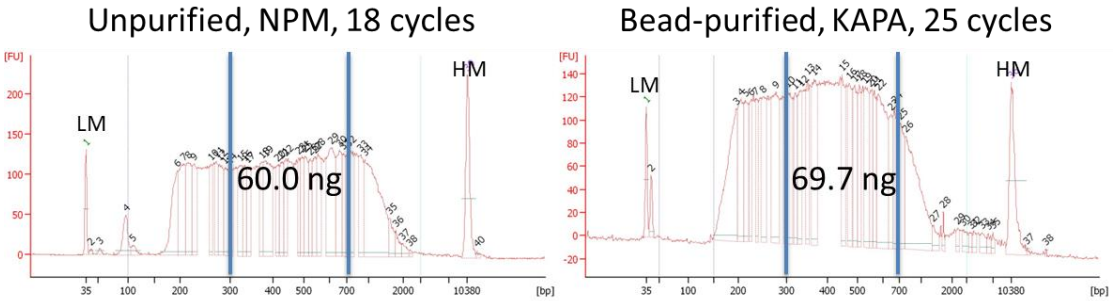


Figure 3.4.3.1 The number of PCR cycles required to amplify non-purified and purified Hi-C libraries to achieve comparable yields

Electropherograms of PCR amplified libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer, showing size distribution patterns of libraries processed under different conditions as labelled above. The area between 300 and 700 bp is marked by blue lines to indicate the range of fragments required for both optimal sequencing and genome mapping of Hi-C contacts. The yield in nanograms (ng) is the calculated yield between 300 and 700 bp of the corresponding library. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. (Both samples used 2.5 pg population Hi-C DNA. The left sample was not bound to beads whereas the right sample was, and the left sample was not trimmed whereas the right sample was trimmed before tagmentation. Both samples were tagmented using 1/100 ATM in half-volume reaction.)

3.4.4. Splitting the PCR into two consecutive reactions improved fragment size distribution

I noticed that when processing single nuclei amplified with 25 cycles of KAPA PCR, extended “tails” of both smaller (<300 bp) and larger (>2000 bp) fragments could be observed in the electropherograms of the amplified libraries (see Figure 3.4.4.1 a for example). These tails may represent either real Hi-C ligated fragments that were over-amplified or instead are signs of biased or abnormal amplification of fragments (ie. single-stranded DNA), which may reduce library quality. One possible cause of this issue was that during the 25 cycles of PCR the reagents such as dNTP’s, primers and polymerase were being exhausted resulting in these abnormal fragments. To investigate this I decided to test whether splitting the reaction into two consecutive reactions, with an extra clean-up step between the two, improved the profile of library fragments. This was done by initially amplifying the library with 9 cycles, purifying the amplicons from this first reaction, then re-amplifying the purified amplicons for a second time with another 16 cycles, before purifying the library again to obtain the final library. When the libraries were analysed using the Bioanalyser, the smaller fragment tail (<300 bp) was almost removed, while the larger sized tail (> 2000 bp) was significantly diminished (see Figure 3.4.4.1 b as an example).

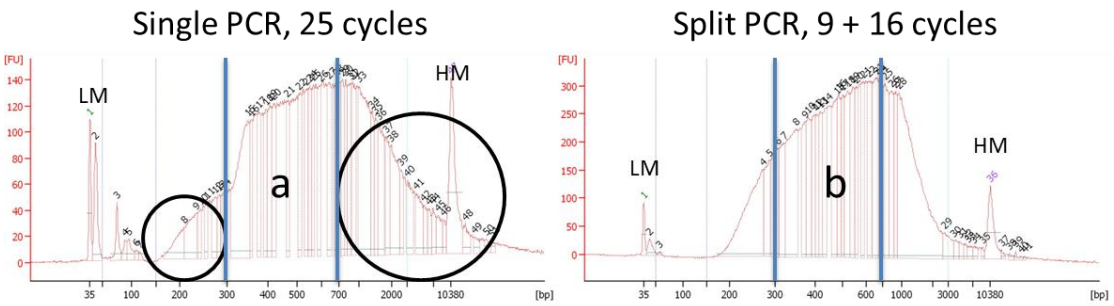


Figure 3.4.4.1 Comparison between single-cell Hi-C libraries amplified by one-round and split PCR

Electropherograms libraries analysed on high-sensitivity DNA chips using an Agilent 2100 Bioanalyzer, showing DNA fragment distribution of libraries amplified by different PCR conditions as labelled above. The circles in sample a indicate the "tails" of abnormal fragments. The area between 300 and 700 bp marked by blue lines indicates the range of fragments required for both optimal sequencing and genome mapping of Hi-C contacts. The axis labels FU and bp indicate the detected amount of DNA in fluorescent units against fragment size in bp. The x axis scale marks from left to right are 35, 50, 100, 150, 200, 300, 400, 500, 600, 700, 1000, 2000, 3000, 5000, 10380 bp. LM and HM indicate the lower and higher molecular weight markers of the chip respectively. (Both libraries were from mouse haploid single nuclei. The samples were bound to streptavidin-coated magnetic beads, trimmed, and tagmented by 1/100 ATM in a half-volume reaction. Amplicons were purified using AMPure XP magnetic beads after each PCR, thus once for sample a and twice in total for sample b.)

3.5. The Transposase method identifies more useful contacts

During the course of my PhD I sequenced around 100 transposase processed single cell Hi-C nuclei. While many did not identify enough Hi-C contacts to be useful for genome structure calculation, one experiment carried out using AluI trimming with 1/100 dilution of transposase and the KAPA PCR system, identified the first two

libraries with sufficient contacts to calculate genome structures[92]. Crucially, this experiment convinced us that developing a single nucleus Hi-C protocol where we could first image the same single cell was possible. Analysis of the contacts from these two datasets revealed that they had a significantly higher "trans ratios" (number of inter-chromosomal contacts over total contacts) over the datasets and structures from the AluI-A-tailing method (see pie chart in Figure 3.5.1).

A contact is a pair of DNA strands that were spatially close to each other in the genome, that were linked by a biotinylated Hi-C junction during processing, and were identified as a valid pair from sequencing reads. Contacts can be classified into intra-chromosomal or cis contacts and inter-chromosomal or trans contacts. In our experience the quality of a single-cell Hi-C structure is largely determined by the number of "useful" Hi-C contacts. A "useful" contact is a contact whose two junction pairs span a minimum of two beads (unified particle representation of certain amount of chromosomal DNA in a bead-on-a-string chromosome computational model) within a calculated polymer model structure, whereas, junction pairs that occur within a structural bead do not provide any useful spatial information for structure determination. For example, if the bead size (or resolution, the amount of DNA in each bead) of a structure is 100 kb then any intra-chromosomal or cis contact that span a distance greater than 100 kb would be termed a "useful" contact. It is natural for some cis contacts being not useful because the sites closer in genomic sequence are more likely to make contacts. In contrast, of course, all trans contacts will be "useful" contacts for structure determination. So a good library aims for higher number of total contacts (trans plus cis) as well as a greater trans component over cis, referred as the "trans ratio". In other words, if two libraries have the same number of total contacts, the library with higher trans ratio will normally result in a better structure.

The trans ratios of good libraries processed by the AluI-A-tailing method normally ranged from 5% to 10%. In contrast, the two libraries processed by the transposase method both had a trans ratio of around 25%. Although in theory higher trans ratio is always better, an abnormally high trans ratio often indicated more than one copy of a

haploid genome had been present in the cell which could not be used in structure calculation due to duplicated genome assignment. To investigate how much better the structure calculated from the transposase library is, another structure was calculated from an AluI-A-tailing library with similar number of total contacts (~30,000) at the same resolution (100 kb) for comparison (Figure 3.5.1). The structure calculation always aims for better resolution or smaller bead size. But smaller bead size will result in more beads overall; each bead will be more likely to be restrained by no experimental data and the structure will be less reliable. 100 kb resolution was chosen because it was the finest resolution that a library with ~30,000 contacts could reach, for both library preparation methods. At this resolution, given that a nucleus has a diameter of approximately 10 μm, the physical radius of a 100 kb bead is roughly 200 nm. As can be seen from the structures, the five models in the structure calculated from the transposase data are more consistent with each other, forming more constrained bundles of DNA fibre. Some of the regions in the AluI-A-tailing genome structure are even not well defined. The difference in structure quality was quantified into the mean pairwise all-particle RMSDs (root-mean-square deviations), which showed a significant improvement from 1.59 to 0.89 particle radii. The RMSD comparison was processed by Dr Tim Stevens. During the process, for each structure five 100 kb models were selected and compared pairwise. The two compared models were first superimposed; then the coordinate variation for each equivalent particle was calculated as particle RMSD; the model RMSD was calculated as the mean value of particle RMSD. Pairwise model RMSDs for all five models were then averaged as the structure RMSD[92].

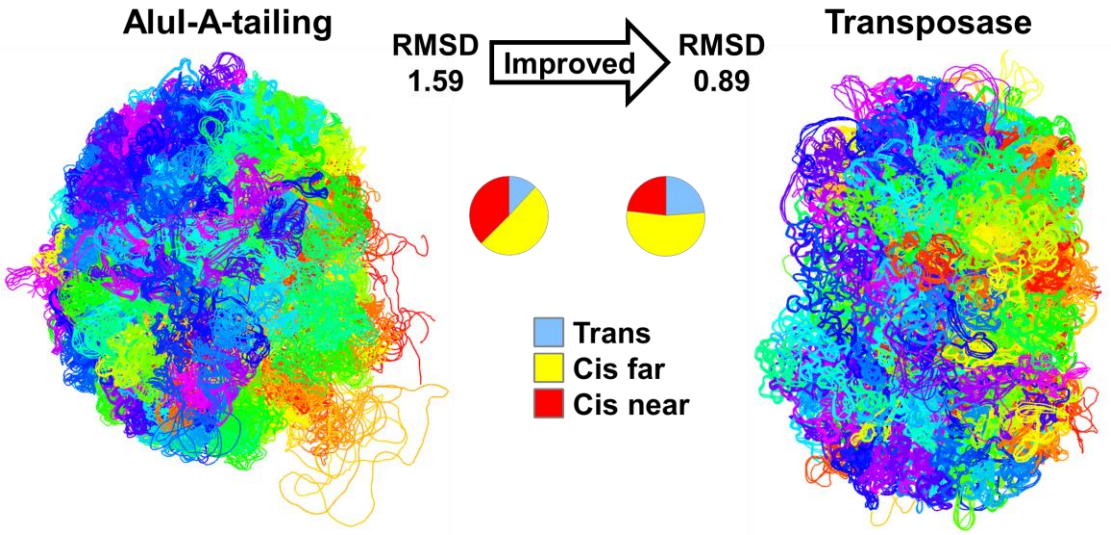


Figure 3.5.1 Improved whole genome structures calculated from transposase processed haploid mouse ES cells

Genome structures calculated at 100 kb resolution using ~30,000 Hi-C contacts identified using AluI-A-tailing (left) and transposase methods (right)[92]. Each structure is an ensemble of five superimposed conformations, from repeat calculations using the same experimental data. Chromosomal positions have been coloured from centromere (red) to telomere (purple). The precision of the structure ensembles is presented in terms of mean pairwise all-particle RMSDs, in units of particle radii. Pie charts show the distributions of Hi-C contacts (trans, cis far (>10kb) and cis near (<10kb)) for each data set. Data analysis and structure calculations were done by Dr Tim Stevens.

However, the AluI-A-tailing method has been considerably improved since this transposase genome structure, whereas I have not been able to produce another good library using the transposase method. The current AluI-A-tailing libraries can have numbers of total contacts as high as 200,000, which are significantly more than the 30,000 contacts of the transposase libraries, resulting in structures with higher resolution despite the lower trans ratios (see Chapter 4 and 5). Also, the number of transposase libraries (only 2) is far from enough to suggest that transposase libraries if prepared in the future would consistently have a trans ratio as high as 25%. In fact, libraries in another similar single-cell Hi-C study using the transposase approach have

trans ratios similar to the numbers of our AluI-A-tailing libraries[67,94]. This may indicate that the high trans ratios of our transposase libraries were not due to the efficient tagmentation in sequencing library preparation, but actually due to an outstanding Hi-C reaction efficiency in that particular experiment. However, a more systematic analysis on sequencing read quality shows that the transposase method has the potential to give contact maps that have less noise because it is a one-step protocol for adding adaptors[94].

3.6. Summary of experiments by transposase

Both Hi-C processed single nuclei and various amounts of population Hi-C DNA were used to test different conditions. After library preparation only the libraries that had good yield and fragment distribution were sent for sequencing. Then their quality was analysed based on the sequencing results. Although both real nuclei and population DNA samples gave high quality results, only good datasets from real nuclei can of course be used to calculate genome structures. Overall, approximately 1200 samples were processed in more than 60 experiments. Modifications that had improved the results in one experiment were inherited to the following experiments. The accepted modifications so far are summarised as follows. LMP agarose left from previous steps is not compatible with the transposase enzymatic reaction, tagmentation, thus should be removed before tagmentation by binding biotinylated DNA to streptavidin magnetic beads. However, in this case, tagmentation has to be carried out on bead-bound DNA, where the original parameters suggested in the kit are again not compatible. To resolve this problem, bead-bound DNA should be trimmed first by AluI restriction. Also the transposase must be diluted to 1/100 of the suggested concentration to avoid over-tagmentation on the single-cell amount DNA. In addition, the PCR system provided in the kit was also found not compatible with tagmented on-bead single-cell Hi-C DNA. Thus the KAPA HiFi PCR polymerase was used instead to consistently amplify the libraries, where two consecutive reactions

with 25 cycles in total were found optimal.

The modified method has successfully produced 2 single-cell Hi-C libraries with approximately 30,000 Hi-C contacts. Although one of them had a missing chromosome probably due to haploid cell cycling, thus could not be used for structure calculation, data from the other library allowed successful genome structure calculations at 100 kb resolution. Its quality was comparable to the best ones processed using the traditional AluI-A-tailing method; and it was used as one (Cell 3) of the eight core datasets in the published paper of my lab[92] (see Chapter 4). However, since these two only successful libraries, all other cells could not give libraries with more than 10,000 contacts with decent quality, thus could not be used for structure calculation. During this period, as well a lot of works were done to troubleshoot the inconsistency. Compare with the protocol used for the successful libraries, no significant further optimisations had been made to the library preparation steps. However improvements had been made on the single-cell sample preparation steps, like the introduction of FACS (Fluorescence-activated cell sorting) instead of hand-picking to isolate single cells, and on Hi-C reaction steps, including optimisations on the concentration of reagents used. Also, based on practice, technically the sample handling in some steps, for example the bead washes, had been significantly improved. Taking advantage of these improvements, the protocol could consistently process 2.5 pg population Hi-C control samples to give libraries with good fragment size distribution profiles. Some of these libraries were sequenced using Illumina MiSeq. Data analysis on the resultant reads further confirmed the sequence quality of these libraries (data not shown). However, single-cell samples processed alongside these control samples still could not give comparable libraries. Further work is needed to understand the reasons for this.

4. 3-Dimensional genome structure of mouse ES cells

Since my group's first publication on single-cell Hi-C in 2013[56], we have continued studying nuclear architecture by further developing the approach. I joined the group for my PhD in 2014 and have been working on the single-cell Hi-C project since then. During this period, the focus of our work was to carry out single cell Hi-C on haploid mouse embryonic stem cells (mESC), so that the first 3-dimensional (3D) structures of an intact whole genome could be determined and studied for the first time. The results were very significant and most findings were published in Stevens et al. in 2017[92]. (In the later context of this chapter, the term "paper" refers to this publication unless specifically indicated.) I contributed the data of one cell out of the eight in the paper, but I was not involved in their analyses. However I did the analysis on the genome and chromosome flatness, which has not been published. In this chapter, I will re-discuss the main findings in the paper (in Sections 4.1 – 4.3 and 4.5 – 4.7) and summarize my further work since the publication (Section 4.4). (Many points of this chapter refer to Stevens et al.[92], and for clarity, repeat citations to this paper will be largely omitted.)

During the cell cycle, the nuclear architecture of mESCs undergoes dramatic changes due to the alternation between DNA synthesis and mitosis (see section 1.4.2 for a detailed introduction). Some features of genome structure such as TADs and A/B compartments disappear in M phase and reform as cells progress into G_1. During G_1 phase, the structures are thought to be in a relatively stable conformation with features required for transcription and cell growth, and these features are believed to be cell type specific. G_1 phase is also the only phase in the cell cycle (except G_0) where one copy of the genome is present. By using haploid mESCs in G_1 it was possible to uniquely map single-cell Hi-C fragment sequences to the correct chromosome, unlike diploid mammalian cells which always contain at least two copies of each chromosome.

Our study was mainly based on data generated from single-cell Hi-C. In addition, we also used single-cell imaging of the same nuclei and we combined the structures with population-based chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) and RNA sequencing (RNA-Seq) data to investigate nuclear organisation.

*Full names of the abbreviation used in this chapter (not included in the List of Abbreviations of this thesis): Dr. Srinjan Basu (SB), Dr. David Lando (DL), Dr. Tim J Stevens (TJS), Dr. Wayne Boucher (WB), Dr. Xiaoyan Ma (XM), Dr. Liam P Atkinson (LPA), Dr. Kai Wohlfahrt (KW) and Prof. Ernest D Laue (EDL).The genome structure studies described here are a highly collaborative effort carried out by the E.D.L. group with the help of several other groups. For the single-cell Hi-C experiments, S.B. and D.L. prepared haploid mESCs and isolated individual nuclei. S.B., D.L. and I carried out imaging. D.L. and S.B. and I processed the Hi-C reactions. D.L. did most of the sequencing library preparation by the AluI-A-tailing method. I processed some cells by the AluI-A-tailing method and carried out all the sequencing library preparation for the transposase method. D.L. and I did the quality control and sample preparation for high-throughput sequencing. T.J.S. and W.B. developed the software for processing the sequencing reads and helped D.L. and I process the data. T.J.S. was the main developer of the structure remodelling software. T.J.S., W.B. and X.M. carried out the structure calculations. T.J.S. and L.P.A. did most of the computational analysis of the structures, and K.W. developed the software for image/structure superposition. E.D.L. conceived the experiments, helped with data analysis and wrote the manuscript with help from the other authors. Other collaborators and their contributions are all listed in the “author contribution” section of the paper[92].

4.1. Consistent single genome structures at 100 kb resolution

*It was mainly me who processed using the transposase method and provided the data of one of the eight cells (Cell 3) discussed in this section. It was mainly D.L. who processed using the AluI-A-tailing method and provided the data of the rest of the cells. It was mainly T.J.S. who did the analyses for this section.

From our single-cell Hi-C experiments on mESCs, we successfully obtained about 20 good datasets of individual cells. Eight of these cells referred to as Cells 1-8 were used in the analysis for the paper (Table 4.1.1), while the rest were mostly obtained after publication. Each cell yielded a minimum of 37,000 Hi-C contacts, corresponding to at least 1.2% recovery of the theoretical ligation junctions. These contacts were classified into intrachromosomal (cis) and interchromosomal (trans) contacts. Trans contacts and cis contacts with a sequence separation greater than 100,000 kb contributed more to the structures calculated to 100 kb resolution, whereas short range cis (< 100,000 kb) contacts were less important. The ratio of trans contacts to total contacts was an important factor contributing to single-cell Hi-C data quality, because a sufficient number of trans contacts is needed to successfully compute an intact genome structure. In population Hi-C carried out on millions of cells, trans contacts are found to be abundant between any two pairs of chromosomes (Figure 4.1.1 a, below the diagonal). In contrast, our single-cell Hi-C contact maps shows clear clusters of trans contacts between certain pairs of chromosome for each cell (Figure 4.1.1 a, above the diagonal). For each specific chromosome, both the pairing and number of contacting chromosomes varied from cell to cell (Figure 4.1.1 c).

Table 4.1.1 Sequencing read data for the 8 published cells

Cell	Input read pairs	Unique mapped pairs	Primary contacts	Final contacts [a]	Normal ligation %	Single read %	Trans %	Promiscuous ends	Mean redundancy
1	1,969,076	1,235,949	110,042	122,475	89.36	21.5	11.65	4342	7.88
2	1,621,648	944,140	65,636	84,129	92.29	21.7	6.30	2320	7.70
3 [b]	1,937,061	1,326,834	32,534	37,604	68.67	14.0	22.66	1826	24.15
4	1,517,614	704,831	75,740	92,748	89.99	23.7	6.64	771	6.39
5	1,592,161	883,678	61,855	74,417	89.86	21.2	9.67	1833	10.12
6	1,493,430	643,721	60,334	75,352	92.21	21.7	5.21	699	7.70
7 [c]	4,796,232	1,191,086	46,438	55,792	80.60	16.5	5.51	1655	17.26
8 [c]	1,776,396	810,661	35,157	42,586	90.40	20.13	7.42	1423	16.00

[a] The final contacts are unique contacts either distinctly mapped to the reference genome or ambiguously mapped but resolved using the 100 kb structures.

[b] Cell processed using the transposase method, while the remaining libraries was obtained using the AluI-A-tailing method.

[c] Cell with structures validated by their corresponding imaging data of centromere positions.

This Table is reproduced from Stevens et al.[92].

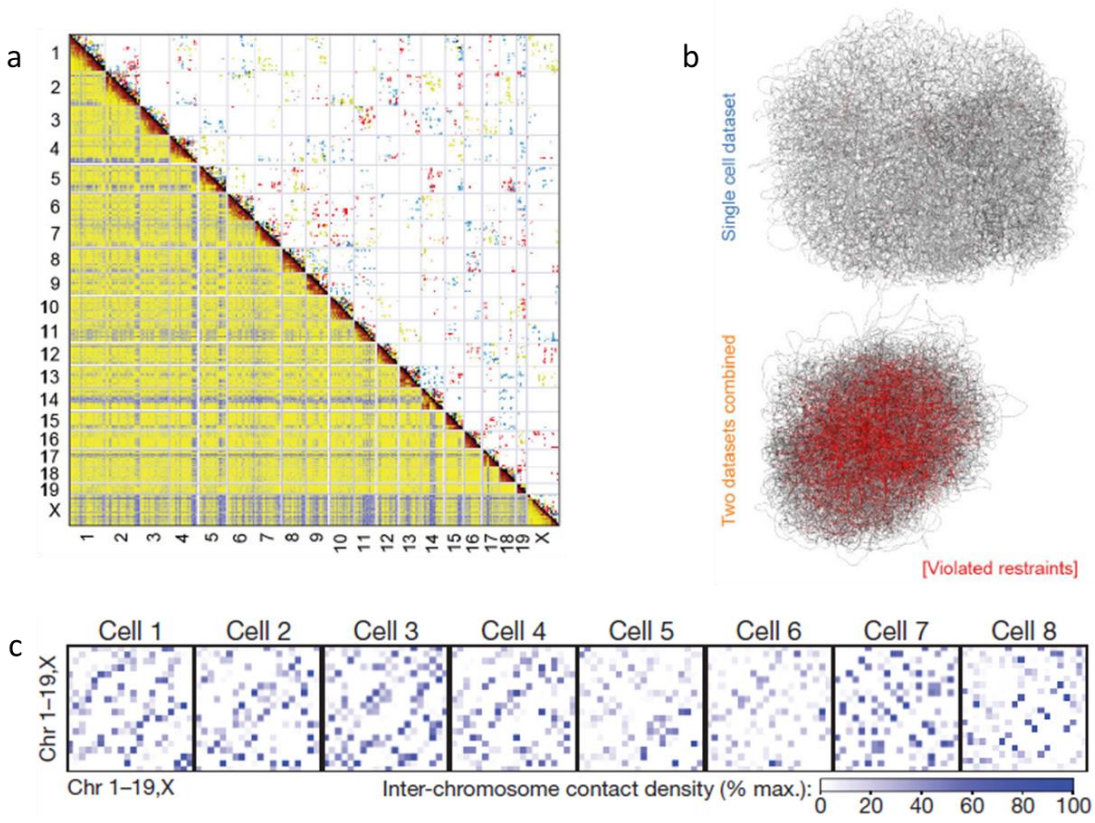


Figure 4.1.1 Trans contacts in single-cell Hi-C maps

a) Comparison between trans contacts of single-cell Hi-C (above the diagonal line) and population Hi-C (below the diagonal line). Data from three different single cells coloured in red, yellow and blue were superimposed. b) Comparison between structures calculated from a

single cell dataset or a merged from two cells' datasets. Strongly violated restraints (distance between the restrained bead pair greater than 4 bead radii) are shown in red. c) Trans contact density between different pairs of chromosomes of the 8 published cells. These figures are reproduced from Stevens et al.[92].

We calculated 3D genome structures for every cell based on its Hi-C contacts using a bead-on-a-string model and an extended simulated annealing protocol. For each cell, we first calculated structures using beads containing 2 Mb DNA, or at 2 Mb resolution. The calculation was repeated 20 times independently to generate 20 models. Then we increased the resolution by reducing the bead size down to 400 kb, then 200 kb, and finally 100 kb and in some cases for the better datasets, 25 kb. At the 400 kb stage, we selected the best 10 models from the 20, as an ensemble with the lowest median root mean square deviation (RMSD) between each pair of models in the ensemble, for further calculations. During the selection, for each ensemble 10 models were compared pairwise. The two compared models were first superimposed; then the coordinate variation for each equivalent particle was calculated as particle RMSD; the model RMSD was calculated as the mean value of particle RMSD. Pairwise model RMSDs for all 10 models were then averaged as the ensemble RMSD[92]. Smaller ensemble RMSD values would indicate higher consistency between the models and a more reliable genome structure at a specific resolution. The RMSD also normally increases at higher resolution. The threshold indicating highly reliable genome structures was set to be less than 1.75 bead radii (unit of distance used in the structure calculation algorithm). With more than 30,000 Hi-C contacts, every cell managed to give a highly reliable structure at 100 kb resolution. A few of the best cells, like cell 1, could be calculated at 25 kb resolution. They normally had more than 100,000 total contacts, a relatively high interchromosomal (trans) contact ratio with little isolated contact noise, such that most contacts had a supporting contact within a 1 to 2 Mb window.

For the same cell, the basic 3D folding conformation of the genome was found not to change, when calculated with randomly selected contacts (down to 30% of total

contacts) were used (Figure 4.1.2). However, the highest resolution with consistent structure, or RMSDs at a particular resolution did change according to the number of contacts (Figure 4.1.2 RMSDs). This type of analysis indicated that a minimum of 10,000 contacts are needed to determine the basic organisation of the chromosome, while more contacts will add details to the underlying folding conformation of each chromosome.

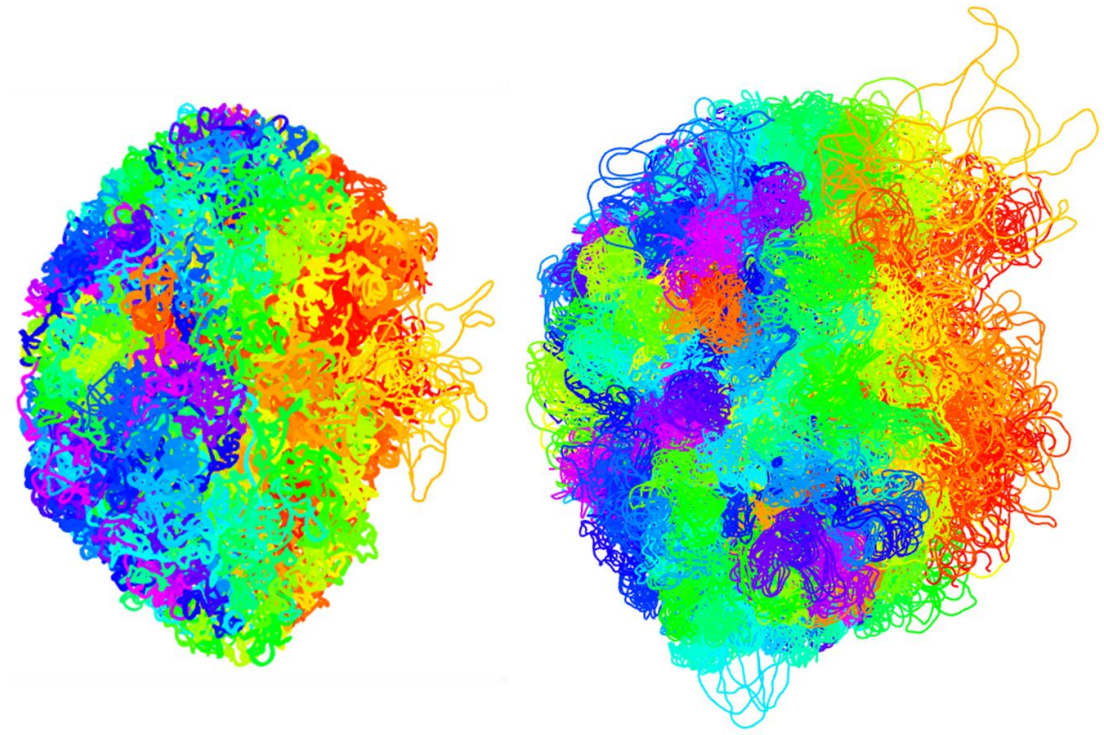


Figure 4.1.2 The same folding conformation calculated from partial contacts

100 kb resolution structures of Cell 1, calculated from the total 122,475 contacts (left) and 20,000 randomly selected contacts (right), are coloured from centromere side (red) to telomere side (purple). Both structures contain 5 aligned models. The average RMSD of the 5 models of each structure is shown at the bottom.

We found that each cell had a genome structure distinct from all the other cells. After randomly merging half of the data from two different cells, the ratio of strongly violated restraints (distance between the restrained bead pair greater than 4 bead radii)

increased from the typical 5-6% to 37.4%. This suggests that genome structures from individual cells are unique. The merged structure derived from this calculation was also highly inconsistent forming a much more condensed structure (Figure 4.1.1 b). In addition, we were able to use these violations to identify any cells with broken, recombined or duplicated chromosomes.

4.2. A Rabl configuration of the chromosomes was revealed by the structures and validated by imaging.

*It was mainly me who processed using the transposase method and provided the data of one of the eight cells (Cell 3) discussed in this section. It was mainly D.L. who processed using the AluI-A-tailing method and provided the data of the rest of the cells. It was mainly T.J.S. and K.W. who did the analyses for this section.

We imaged the centromere protein CENP-A to validate our structures (Figure 4.2.1 a). It should be noted that our structures do not include the actual centromeres/telomeres because contacts cannot be mapped to the repetitive sequences found at centromeres and telomeres. Thus the centromere/telomere positions in our structures are defined as the positions of the nearest mappable sequence on the chromosome. The size of a typical mouse centromere is approximately 300 kb[95,96]. So, considering the centre of this region, the predicted position is about 150 kb or 1 – 2 beads of 100 kb away from the actual position. The actual centromere positions were determined by imaging fluorescently labelled CENP-A (the centromeric histone H3 variant) expressed in one of the same cells that were later processed for single-cell Hi-C. We then validated our structures by superpositioning the centromere clusters derived from the CENP-A image with the positions of the centromeres in the calculated genome structures. In the paper we imaged two of the 8 cells (Cells 7 and 8), and both showed good correlation between the two types of centromere positions providing evidence that the calculated genome structures are correct (Figure 4.2.1 b).

From the modelled structures of all G_1-phase haploid mESCs and the CENP-A images,

we observed a trend of centromeres and telomeres clustering on opposite sides of the cell (Figure 4.2.1 b, c). This is consistent with the Rabl configuration, first described by Carl Rabl in the late 19th century, where centromeres of interphase chromosomes where found to cluster on one side of the cell. Such consistency from cell to cell also strongly validates our structures. For example, centromeres of cell 7 are clearly clustered in a cavity on one side of the structure; a few centromeres of cell 8 were more diffused, but the majority still clustered in a similar way to cell 7 (Figure 4.2.1 b, c).

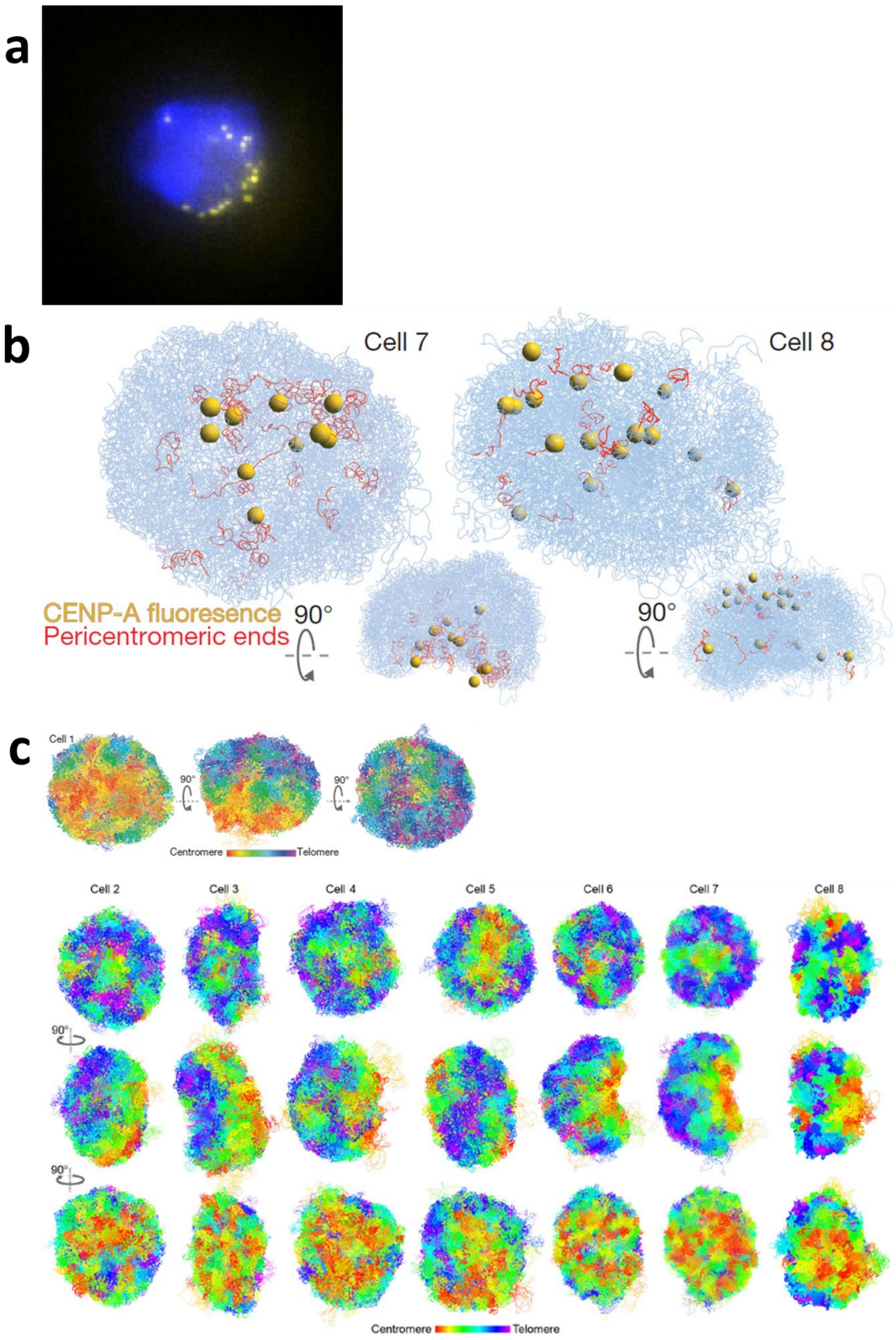


Figure 4.2.1 Rabl configuration and centromere clustering

a) An example of CENP-A (yellow dots) and H2B (blue) image on a haploid G1-phase mESC. b) 3D genome structures of haploid mESCs with corresponding positions of fluorescently imaged CENP-A protein. The centromere ends of the chromosomes are coloured in red.

CENP-A clusters detected from microscopy imaging are shown as yellow spheres. Views of the structures rotated through a 90 ° angle are also shown. Both cell 7 and cell 8 show clustering of centromeres on one side of the genome. c) 3D genome structures of haploid mESCs showing a Rabl configuration, viewed from three perpendicular angles. Each chromosome is coloured from red (centromere side) to purple (telomere side). These figures are reproduced from Stevens et al.[92].

4.3. Discrete chromosome territories with unique shapes

*It was mainly me who processed using the transposase method and provided the data of one of the eight cells (Cell 3) discussed in this section. It was mainly D.L. who processed using the AluI-A-tailing method and provided the data of the rest of the cells. It was mainly T.J.S. who did the analyses for this section.

In our haploid G_1-phase mESC genome structures, every chromosome forms its own territory mostly discrete from other territories in the same cell (Figure 4.3.1 a, b). This feature is generally consistent in all cells. Typically, only 5-10% of each territory is intermingled with the others (Figure 4.3.1 c).

However, an interesting observation was that the folding of each individual chromosome varies remarkably from cell to cell (Figure 4.3.1 d). In some cells a chromosome can be found to be compact while in other cells it can be found to be in an extended conformation. This variation from cell to cell is consistent with our previous finding that individual chromosomes have distinct trans contact profiles in different cells (see Section 4.1 and Figure 4.1.1 c).

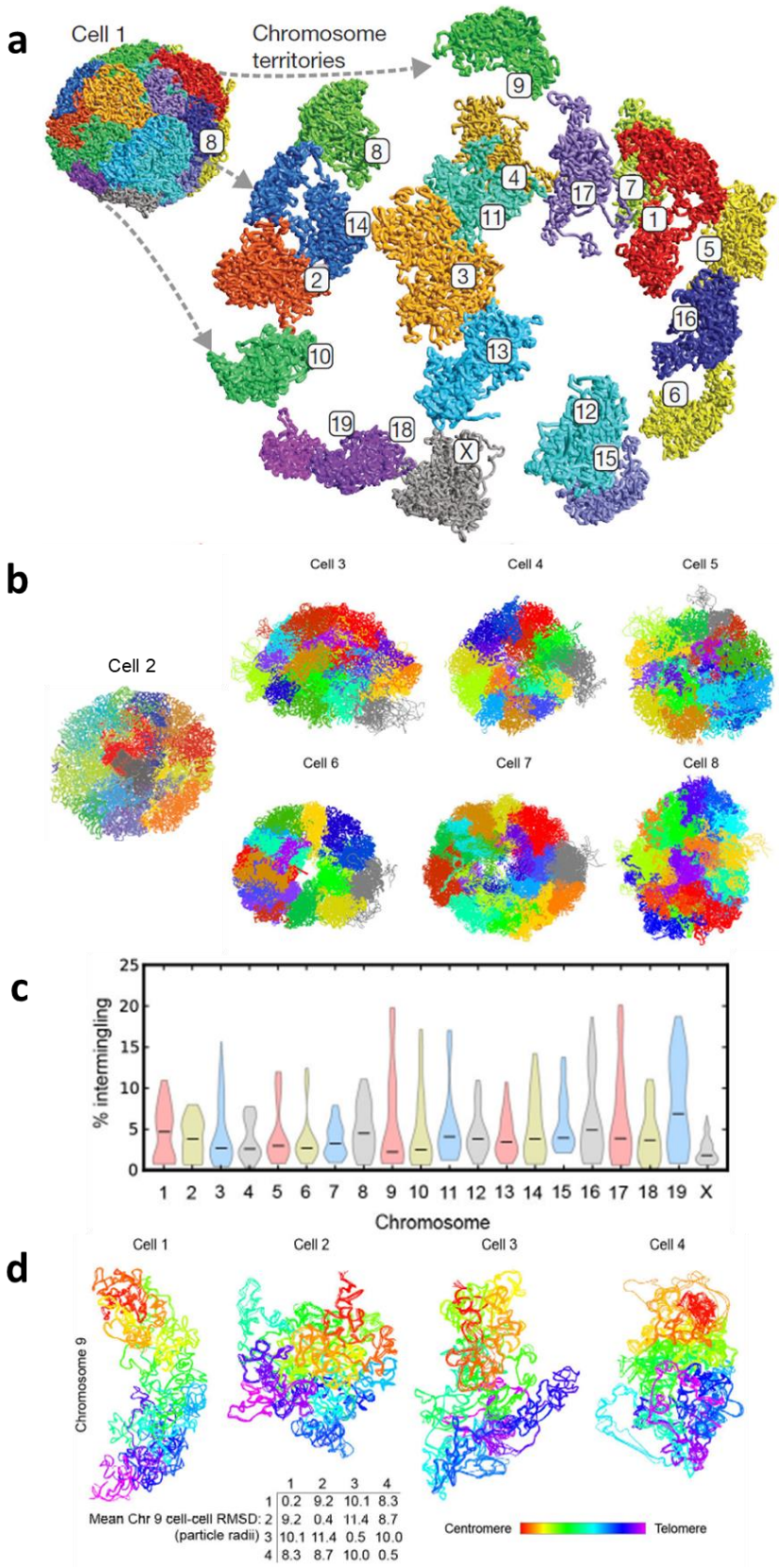


Figure 4.3.1 Chromosome territories of haploid mESCs

a) 3D genome structure of a haploid mESC with chromosomes territories 1-19 coloured from red to purple. The folding and relative positions of individual chromosomes are shown in the

expanded view. b) 3D genome structure of other haploid mESCs showing consistently discrete chromosome territories. c) Violin plot of chromosome intermingling for each numbered chromosome in Cells 1-8. d) The folding of individual chromosomes as compared by pairwise RMSD analysis. Four models of chromosome 9 in Cells 1-4 are shown and coloured from the centromere end (red) to the telomere end (purple). The table below shows the pairwise RMSD values of chromosomes for each pair of cells. These figures are reproduced from Stevens et al.[92].

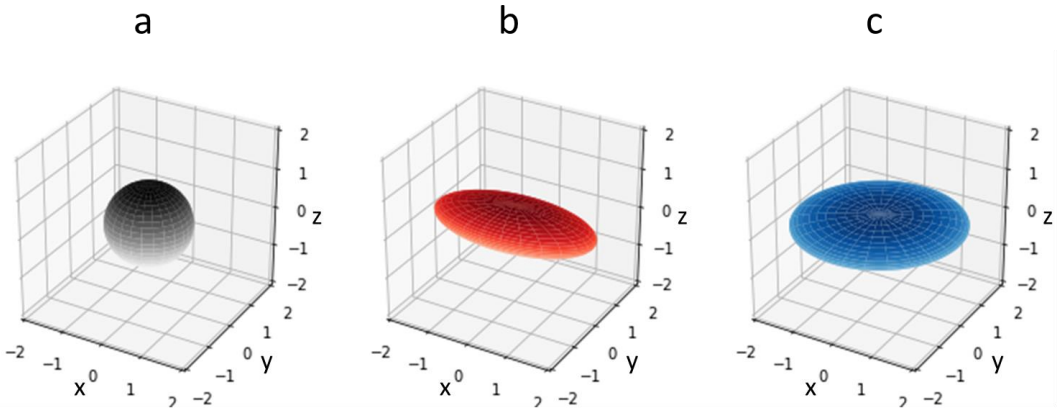
4.4. Relatively consistent genome flatness compared with chromosomes

*It was mainly D.L. who processed using the AluI-A-tailing method and provided the data of the 11 cells discussed in this section. It was mainly me who did the analyses for this section. W.B. helped develop the computational codes for moment of inertia calculation.

I discussed in Section 4.3 that chromosomes from mESC in different haploid G_1-phase have a distinct shape. However it is noticeable in the calculated structures that the whole intact genome shape is approximately an ellipsoid, which is consistent across all the mESC structures.

To mathematically and systematically study the shape of chromosomes within a genome, we developed a strategy using a physical term, moment of inertia (I). The term describes the torque needed for a desired angular acceleration about a rotational axis to rotate a rigid body. In other words, the bigger the value of I, the harder it rotates about the axis, the more flattened the rigid body shape is on the orthogonal plane of the axis. A rigid body would have three I values (I_x, I_y and I_z) along the three orthogonal axes (x, y and z) in 3D space, and the pairwise ratio of them represents to what extent the shape extends regardless of size. For examples, a sphere would have the same I_x, I_y and I_z and all pairwise ratios of 1 (Figure 4.4.1 a); a rod would have a small I_x along the axis of length and big I_y, and I_z along the axes of diameter, with

I_z/I_y=1 and I_z/I_x and I_y/I_x greater than 1 (Figure 4.4.1 b); and a plate would have a big I_z along the axis of height and relatively small I_x and I_y along the axes of diameter, with I_x/I_y=1 and I_y/I_x and I_z/I_x greater than 1 (Figure 4.4.1 c). For a more organised representation, in our analysis I_x, I_y and I_z were sorted from the smallest to largest value. By definition, the structure is the most extended along the x axis, then the y axis, and the least extended along the z axis. Then with the smallest I_x, as a reference, only I_y/I_x and I_z/I_x were calculated where I_z/I_x was no smaller than I_y/I_x, and I_y/I_x was no smaller than 1. The square root (sqrt) of both ratios (sqrt(I_y/I_x) and sqrt(I_z/I_x), referred as "*I* ratios") was used for one-dimensional representations and linear scale comparison. In general, the greater absolute values of sqrt(I_y/I_x) and sqrt(I_z/I_x), the more extended the shape is along the x axis (e.g. Figure 4.4.1 b). The greater differences between sqrt(I_y/I_x) and sqrt(I_z/I_x), the more compressed the shape is along the z axis, the more extended along both the x and y axes, or the more flat on the x-y plane (e.g. Figure 4.4.1 c).



r_x	r_y	r_z	r_x	r_y	r_z	r_x	r_y	r_z
1	1	1	2	1	0.5	2	2	0.25
I_x	I_y	I_z	I_x	I_y	I_z	I_x	I_y	I_z
0.4	0.4	0.4	0.25	0.85	1	0.8125	0.8125	1.6
	Sqrt(I_y/I_x)	Sqrt(I_z/I_x)		Sqrt(I_y/I_x)	Sqrt(I_z/I_x)		Sqrt(I_y/I_x)	Sqrt(I_z/I_x)
	1	1		1.844	2		1	1.403

Figure 4.4.1 Simple geometric examples of moment of inertia

a, b and c are example rigid bodies of a sphere, spindle ellipsoid and oblate ellipsoid respectively, of the same volume and in uniform density. The radii (r_x, r_y, r_z), moments of

inertia (I_x, I_y, I_z) and I ratios corresponding to the three principle axes of each ellipsoid are shown in the table below.

I calculated the I ratios of all chromosomes and the whole genome of the published eight cells and three recently processed single-cell Hi-C datasets of mESCs, labelled Cell 9 – 11 (Figure 4.4.2 and 4.4.3). The three I values were calculated from the 10 structure models of each cell, then averaged for I ratio calculation. To simplify the geometric realisation of I ratios, chromosomes and genomes are assumed to be ellipsoids before compared with their real structures.

Figure 4.4.2 a shows that the I ratios of most chromosomes vary significantly from cell to cell, consistent with the results shown in Section 4.3. In particular, chromosomes 13 and 17 show the greatest variation whereas the chromosomes 4 and X show the most concentrated distribution. By comparing the medians and the means, the I ratios of chromosomes 5, 7 and 12 are relatively small while the ratios of chromosomes 3 and 6 are the highest. As exemplified in Cell 10 (Figure 4.4.2 b), I ratios also vary between different chromosomes in each cell. In this particular cell, chromosomes 3 and 19 have the highest I ratios. This indicates that these chromosomes are in a relatively more elongated shape which is verified by their corresponding structures (Figure 4.4.4 a, compare chromosomes 1 and 9 with 3 and 19).

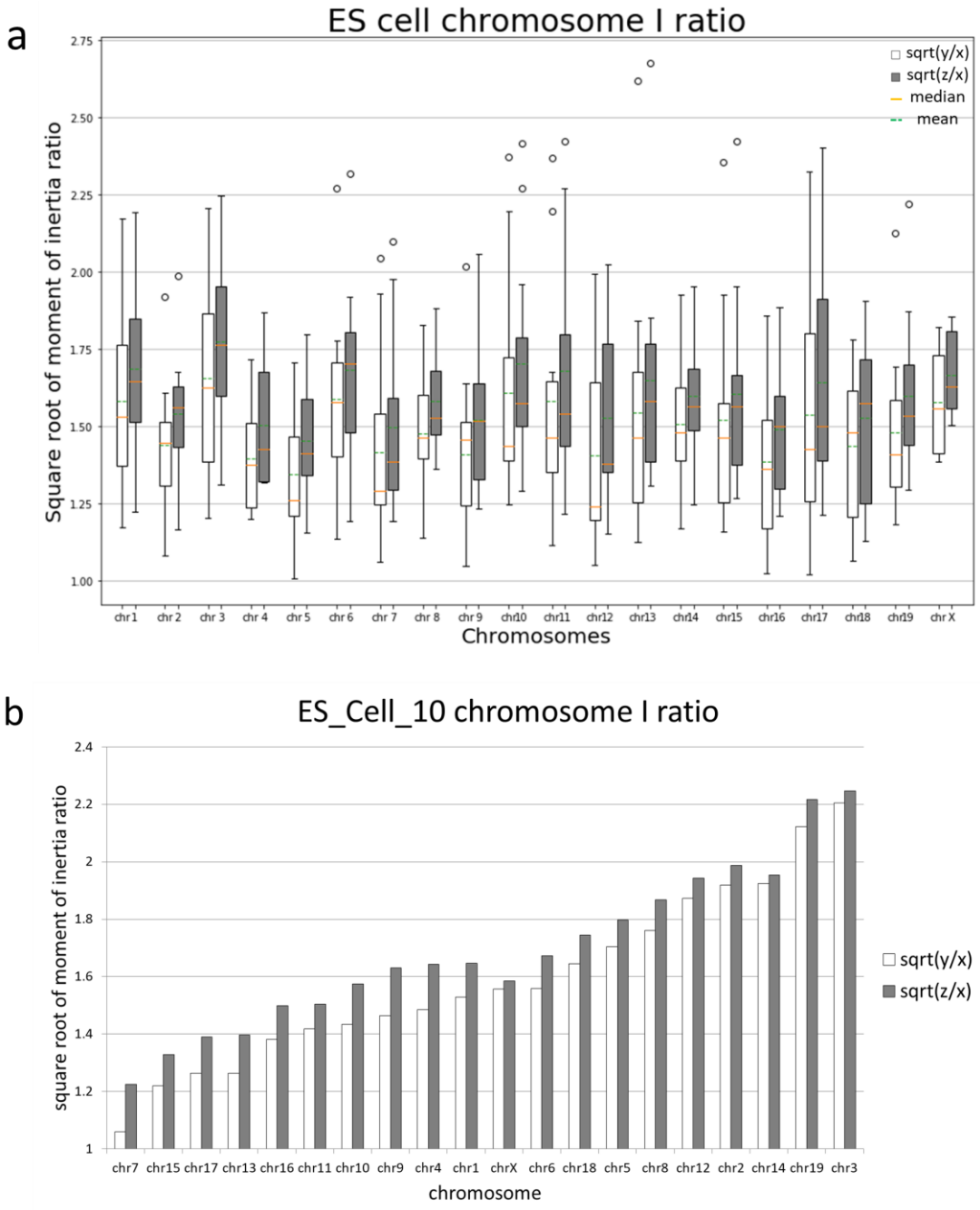


Figure 4.4.2 *I* ratio distribution of ES cell chromosomes

a) Box plot of each chromosome's *I* ratios of the 11 ES cells. White and grey boxes represent sqrt(I_y/I_x) and sqrt(I_z/I_x) respectively. Circles represent the outliers in the corresponding distributions. Within each box, the orange line represents the median and the dashed green line represents the mean. b) Bar chart of the chromosome *I* ratios of ES Cell 10, sorted for sqrt(I_y/I_x) in ascending order. White and grey bars represent sqrt(I_y/I_x) and sqrt(I_z/I_x) respectively.

As shown in figure 4.4.3, among ES cells the *I* ratios of the whole genome varies

from cell to cell, but to a lower extent compared with chromosomes. The highest I ratios, 1.35 and 1.43 for sqrt(I_y/I_x) and sqrt(I_z/I_x) respectively, occur in ES Cell 11. This indicates that it has the most spindle ellipsoid structure among the cells. ES Cell 9 has the greatest difference between sqrt(I_y/I_x) and sqrt(I_z/I_x), with a value of 0.172, which indicates a relatively flattened ellipsoid structure. Both these findings correlate well with the corresponding structures (Figure 4.4.4). In general, all genomes of the analysed ES cells stay in an ellipsoid structure that is not drastically stretched or flattened.

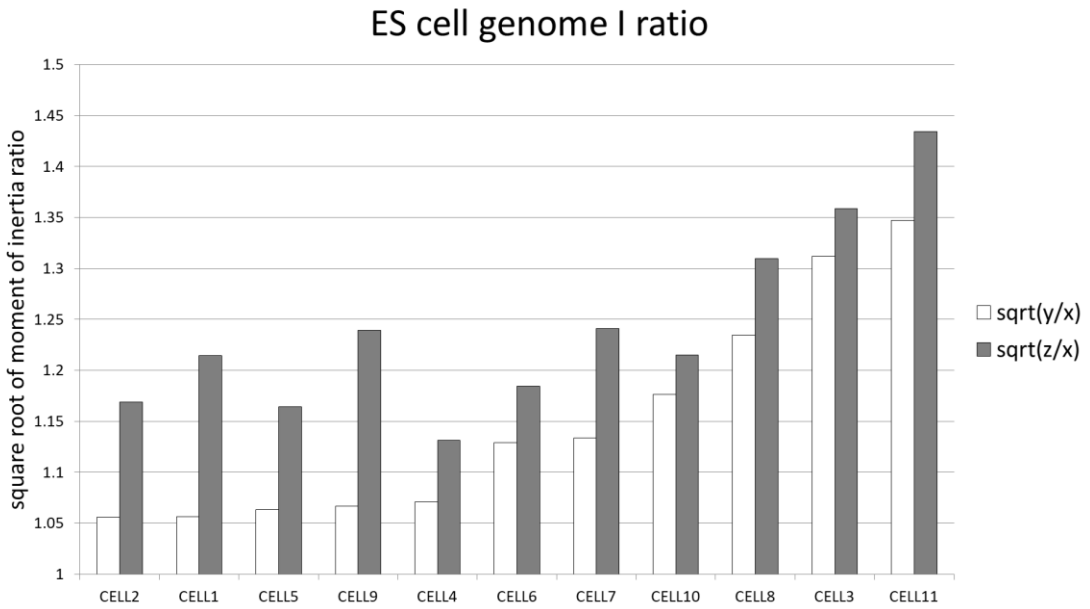


Figure 4.4.3 I ratios of ES cell genomes

Bar chart of the genome *I* ratios of ES cells, sorted for sqrt(I_y/I_x) in ascending order. White and grey bars represent sqrt(I_y/I_x) and sqrt(I_z/I_x) respectively.

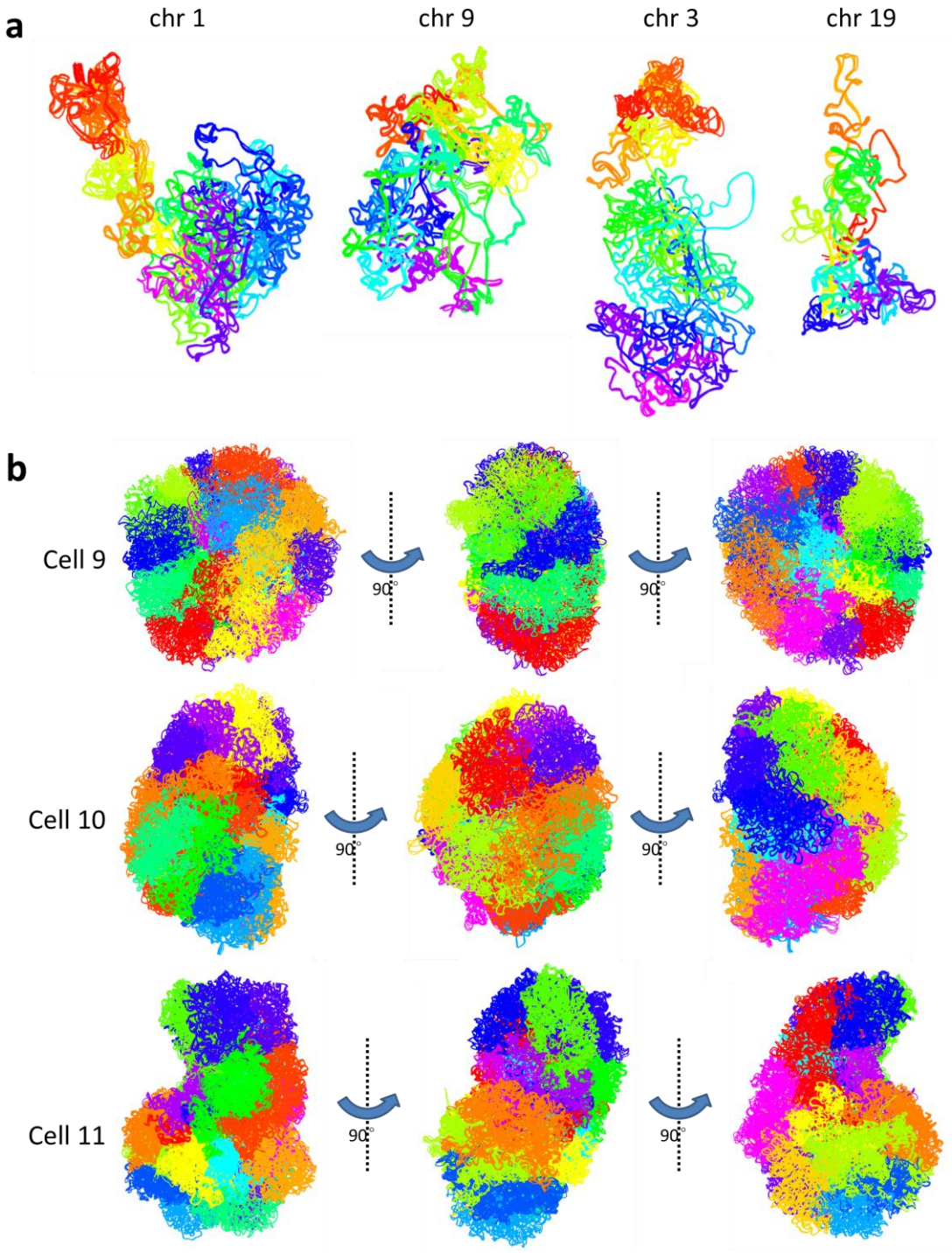


Figure 4.4.4 Structures in ES cells with different flatness

a) Calculated structures of chromosomes 1, 9, 3 and 19 of Cell 10, coloured from centromere side (red) to telomere side (purple). b) Calculated structures of whole genomes of Cell 9, 10 and 11, coloured by chromosomes. 5 overlaid models are shown for each structure in b and c.

4.5. Highly consistent A/B compartments

*It was mainly me who processed using the transposase method and provided the data of one of the eight cells (Cell 3) discussed in this section. It was mainly D.L. who processed using the AluI-A-tailing method and provided the data of the rest of the cells. It was mainly T.J.S. and L.P.A. who did the analyses for this section.

We used population Hi-C data to calculate A/B compartment profiles for haploid mESCs cultured in the same conditions as cells used for single-cell Hi-C. By mapping the compartment data on to our single genome structures, we assigned the structure beads as belonging to either the A or B compartment. When visualizing the A/B identity distribution in a whole genome structure, we find the beads with the same identity tend to cluster together and segregate from beads with the opposite identity. Regions from different chromosomes also cluster or segregate resulting in an outer B compartment shell, a middle A compartment layer, and an internal B compartment shell around a hollow cavity that most likely represents the nucleolus. In our structures, the hollow cavity appears to be contiguous with the nuclear membrane, forcing the A compartment layer into a bowl-like conformation. This B-A-B bowl structure is highly consistent in all haploid mESCs (Figure 4.5.1 a). Chromosome territories occupy different regions of this bowl structure depending on the cell. Chromosomes may be relatively parallel to the curved shape, with a bipolar B-A identity from the outer B shell to the middle A layer (Figure 4.5.1 b middle); may start from the outer B, entering the A, and folding back to the outer B (Figure 4.5.1 b left); or go all the way through the three layers with a B-A-B alternating identity (Figure 4.5.1 b right). This A/B compartment configuration of chromosomes was also shown in a recent imaging study, which helped to validated our structures[97].

By mapping constitutive lamina-associated domain (cLAD)[98,99] data onto our single genome structures, we found these regions always associated with either the nuclear membrane or nucleolar periphery, highly overlapping with the B compartment shell (Figure 4.5.1 a). In contrast, by mapping RNA-Seq data onto the structures, we found that most highly expressed genes are found in the A compartment layer (Figure 4.5.1

a). Similar to A/B compartment profiles, this is consistent in all cells.

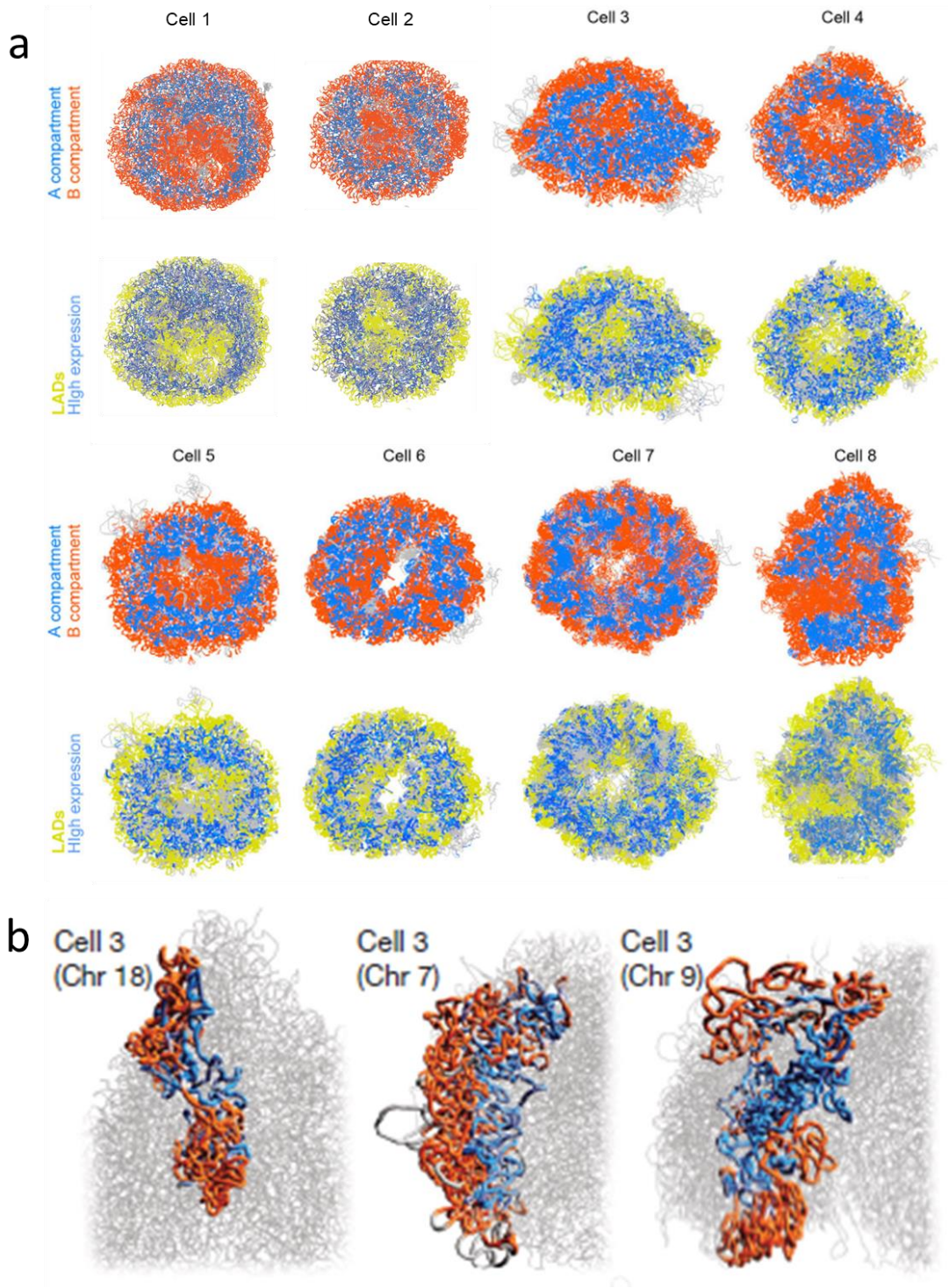


Figure 4.5.1 A/B compartment profiles in single genome structures

a) 3D genome structure cross-sections of the eight published cells. For each cell, the top structure shows the A/B compartments coloured blue and red respectively; the bottom structure shows the cLAD regions (yellow) and highly expressed genes (blue). b) 3D

structures of individual chromosomes highlighted from the whole genome structure. Regions with A compartment identity are coloured in blue and regions with B identity in red. These figures are reproduced from Stevens et al.[92].

4.6. Cell-specific TADs and CTCF/cohesin loops

*It was mainly me who processed and provided the data of one of the eight cells (Cell 3) discussed in this section. It was mainly D.L. who processed and provided the data of the rest of the cells. It was mainly T.J.S. and L.P.A. who did the analyses for this section.

Based on population Hi-C studies, TADs were once thought to be highly compacted and invariant across a cell population[36–38]. However, in our single genome structures, a particular TAD is only compacted in certain cells. In these cells the two TAD boundaries are often close enough to interact, but in other cells where the boundaries are far away from each other, the TAD structure is often rather extended (Figure 4.6.1 a compare cell 1 and 5 with cells 2 and 4). This is consistent with the loop-extrusion mechanism for CTCF/cohesin looping underlying TAD formation and stabilisation[35,49]. Indeed, local loop structures can be seen within TADs in our 3D genome structures consistent with loop extrusion (Figure 4.6.1 a).

By mapping the CTCF/cohesin loop data from mouse B lymphoblasts on to our single genome structures, we found that on average only 62.1% of the loops were present in the published 8 cells (an analysis of selected large loops is shown in Figure 4.6.1 b).

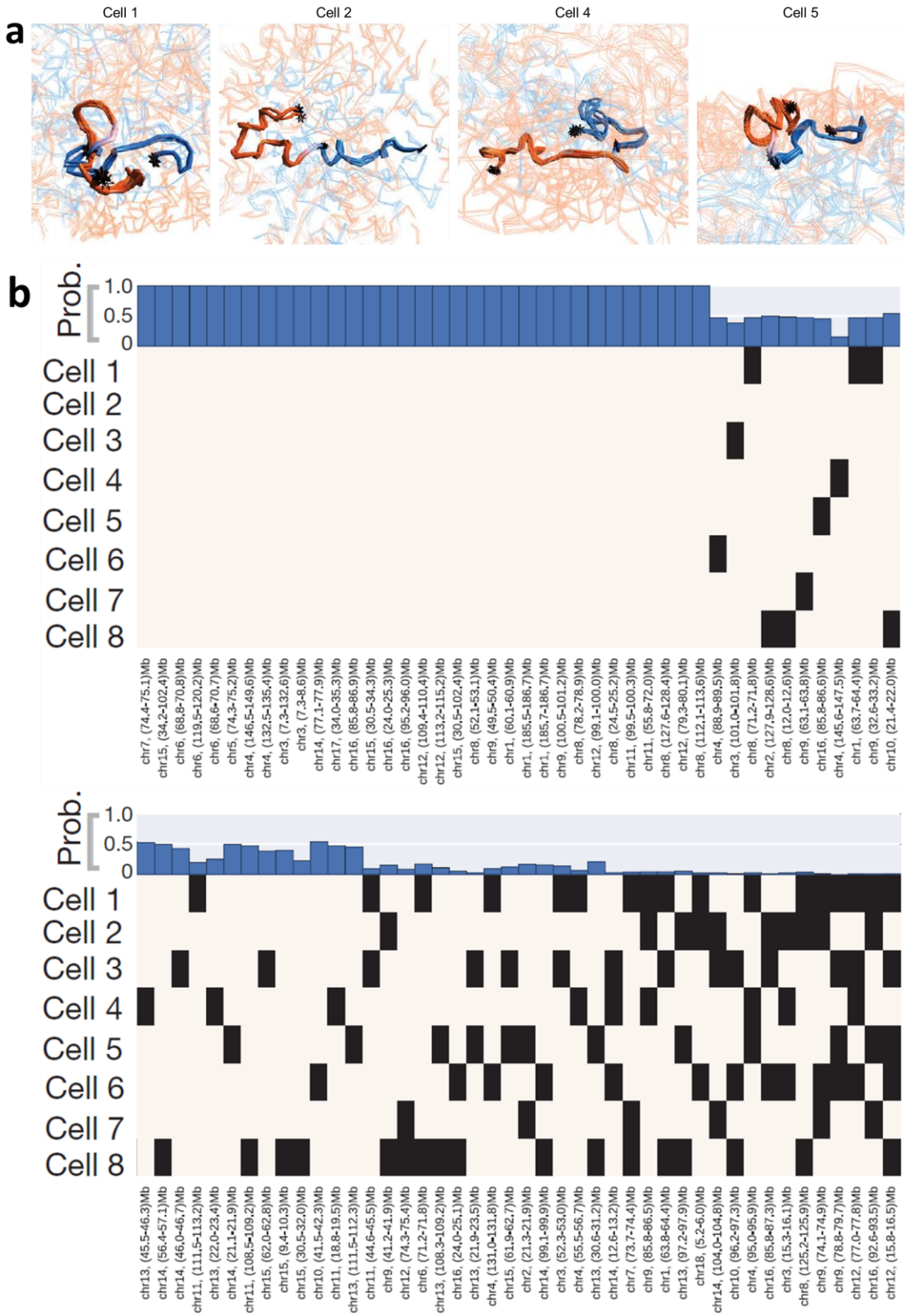


Figure 4.6.1 TADs and CTCF/cohesin loops in single cells

a) 3D structures of a TAD containing region on chromosome 12 from 4 different cells. Structures are comprised of 5 superimposed models from independent calculations. The structures are coloured according to the A/B compartment identity (blue/red). The particular

TAD region in each cell is shown in thicken lines and the two boundaries at sequence positions 67.2 Mb and 70.5 Mb are shown as black asterisks. b) Graph showing whether selected CTCF/cohesin loops could be formed in different single cells. Loops having >600 kb sequence separation based on the data described by Rao et al.[35], and where the two boundaries are close enough to interact are indicated by black squares. Loops with non-interacting boundaries are shown as white squares. These figures are reproduced from Stevens et al.[92].

4.7. Analysis of gene expression and epigenomic features

*It was mainly me who processed and provided the data of one of the eight cells (Cell 3) discussed in this section. It was mainly D.L. who processed and provided the data of the rest of the cells. It was mainly T.J.S. who did the analyses for this section.

We also mapped RNA-Seq and ChIP-Seq data on to our single genome structures to investigate 3D spatial clustering of these features within our structures. Both types of data were obtained from population experiments using mESCs. It should be noted that some of the data was obtained from diploid cells, but for some datasets (such as H3K4me3, H3K27me3 and total mRNA) we verified that the ChIP profile and mRNA expression was similar between haploid and diploid cells.

Based on the ChIP-Seq data, histone H3K4me1, H3K27ac and H3K4me3 were shown to cluster in 3D space in our single genome structures (Figure 4.7.1 a). These euchromatin marks found at active enhancers and promoters also cluster with certain transcription factors such as *Klf4*, which is involved in stem cell identity. In contrast, these euchromatic chromatin marks show little clustering with pluripotency factors such as *Nanog*, or facultative heterochromatin marks such as H3K27me3. Interestingly, all these modified histone H3 marks and transcription factors are anti-correlated to constitutive heterochromatin marks such as histone H3K9me3, which did not show any tendency to cluster in our structures (Figure 4.7.1 a).

To further study clusters of enhancers and promoters, we first categorised activity of

enhancers and promoters, based on combinations of histone H3 modifications present in their genomic regions, derived from ChIP-Seq data. For examples, active enhancers have H3K4me1 and H3K27ac and no H3K4me3, whilst active promoters have H3K4me3 and H3K27ac and no H3K27me3. After annotating genomic regions according to their activities in our structures, we found clear co-localisation between different active enhancers, and active enhancers and active promoters (Figure 4.7.1 b). Active enhancers and promoters are also more likely to be at chromosome interfaces (Figure 4.7.1 d). This is consistent with our finding that highly expressed genes in our RNA-Seq data also prefer to be located at chromosome interfaces (Figure 4.7.1 c left). Highly expressed genes are also likely to be deeper in the A compartment (Figure 4.7.1 c right).

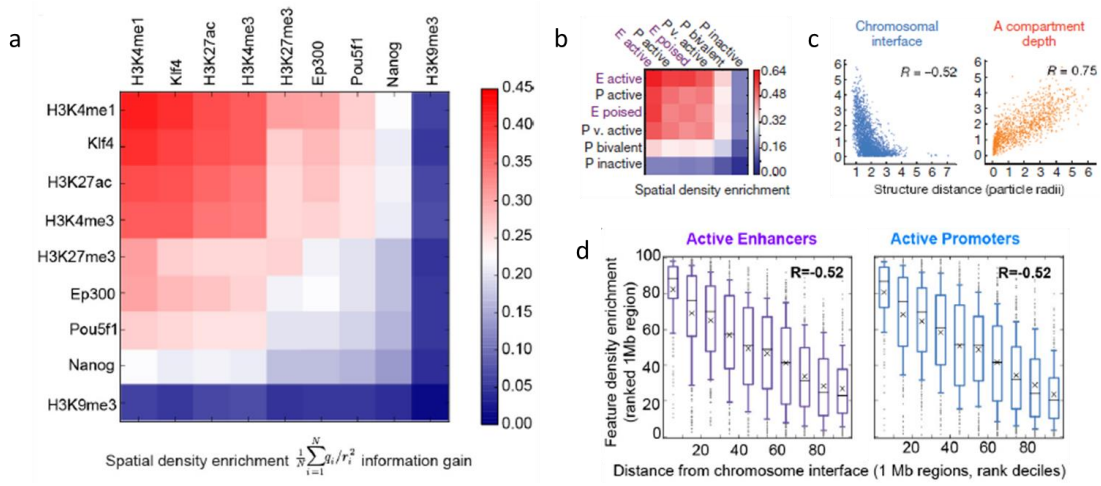


Figure 4.7.1 3D Genome structure and gene expression

a) Pairwise spatial density enrichment of various modified forms of histone H3 and selected transcription factors. b) Pairwise spatial density enrichment of various enhancer (E) and promoter (P) subclasses. c) Plots of spatial density enrichment of gene expression against distance to the closest chromosome interface (left) and A compartment depth (right). Distances are in the unit of bead radii. The R values are the Pearson's correlation coefficient on the underlying, unranked data. d) Density enrichment of active enhancers (purple) and active promoters (blue) at different distance from the closest chromosome interface. Whiskers represent intervals between 10% and 90% of sorted candidates in each group, and boxes represent intervals between 25% and 75%. Crosses represent mean values and bars represent

median values. These figures are reproduced from Stevens et al.[92].

5. 3-Dimensional genome structure of early differentiated cells

It is believed that genome structure in different cell types of an organism is important for distinct functions. My group and I have shown by single-cell Hi-C that certain features of genome organisations in haploid mouse embryonic stem cells (mESCs), such as A/B compartment, LAD and transcription is highly consistent between individual cells (Chapter 4). The next key questions we wished to address was how do all these structural features appear in other cell types; and to what extent are they consistent or differ from ES cells. To try to answer these questions, we decided to carry out single-cell Hi-C experiments on haploid mouse cells in early stages of ES cell differentiation. By studying genome structures at consecutive time points during the differentiation process, it would be possible to monitor their changes from ES cells. The differentiation method also allowed tracking of changes such as those in gene expression levels of key pluripotency factors (Section 1.4.3), which might be related to changes in genome structure.

We designed the experiment to collect cells at two differentiated time points, 24 hours (h) and 48 h (see Section 1.4.3 for detailed methods) and process them by the same single-cell Hi-C protocol as we used for mESCs. For the 24 h time point, we also classified the cells into two groups, high and low *Rex1*-GFP, based on their expression of a GFP reporter gene incorporated at the *Rex1* locus. The *Rex1* gene is a pluripotent marker which has been shown to be highly expressed in pluripotent ES cells. As cells differentiate *Rex1* gene expression is gradually turned off, with bimodal expression at 24h and virtually no expression at 48 h (discussed in Section 1.4.3). Thus, the idea at 24 h was to investigate any differences in genome structure between the two groups of 24 h differentiated cells in a mixed population.

Section 5.1 summarises the collected single cell datasets at the three time points and conditions at the current stage. My group and I are still working on more single-cell Hi-C datasets and other complementary data like A/B compartment by population

Hi-C. Thus at the moment it is not possible for a systematic analysis comparable to the analysis for mESCs as discussed in Chapter 4. However, I have carried out some preliminary analysis on the flatness of chromosome and genome structures which will be discussed in Section 5.2 and 5.3.

5.1. Data collected from differentiated cells have numbers and quality comparable to ES cells

So far, we successfully collected 26 single-cell Hi-C datasets of differentiated cells. All these datasets have over 30,000 contacts with many containing over 70,000 contacts (some of these datasets are shown in Table 5.1.1). They all can be used to calculate consistent single genome structures at 100 kb resolution with some of the better datasets at 25kb resolution. These qualities are comparable to the best datasets of mESCs, indicating that the single-cell Hi-C procedure can be carried out with differentiated cells (see Table 5.1.1 for a comparison). In addition, each condition has at least 6 cells with data collected (6 cells of 24 h differentiated *Rex1*^{high}, 11 cells of 24 h *Rex1*^{low} and 9 cells of 48 h), which indicates that the protocol worked well for all these conditions (compare Table 5.1.1 24 h *Rex1*^{high}, 24 h *Rex1*^{low} and 48 h cells).

Table 5.1.1 Sequencing read analysis of various single cell libraries

Sample[a]	Relative no. of contact[b]	Input read pairs[c]	Uniquely mapped pairs[d]	Deduplicated total contacts[e]
ES Cell 1	max	1969076	1235949	110042
ES Cell 6	med	1493430	643721	60334
ES Cell 3	min	1776396	810661	35157
24 h *Rex1*^{high} Cell 5	max	8167306	4692853	201451
24 h *Rex1*^{high} Cell 6	med	9288453	5061829	111412
24 h *Rex1*^{high} Cell 2	min	9512686	5226588	101843
24 h *Rex1*^{low} Cell 6	max	8764583	5248580	170685
24 h *Rex1*^{low} Cell 11	med	9132954	5228718	123707
24 h *Rex1*^{low} Cell 4	min	6985362	3954004	32782
48 h Cell 2	max	47155006	23990575	134838
48 h Cell 6	med	6632760	3566950	50729
48 h Cell 5	min	7276700	3989993	36819

[a] Three-sample groups comprising differentiated cells and the published mESCs discussed in Chapter 4.

[b] Represent samples with the most, median and fewest "deduplicated total contacts" in the corresponding group.

[c] Total number of paired-end reads for each sample.

[d] Hi-C contact read pairs that map to unique positions in the reference genome.

[e] Total contacts after removing PCR sequencing duplicates.

5.2. Varied chromosome flatness in differentiated cells

Chromosome flatness in differentiated cells was analysed using the same moment of

inertia (*I*) strategy described in Section 4.4. The three *I* values were calculated from the 10 structure models of each cell, then averaged for *I* ratio ((sqrt(I_y/I_x) and sqrt(I_z/I_x)) calculation. The *I* ratios were calculated for all individual chromosomes, for the 6 cells of 24 h *Rex1*^{high}, the 11 cells of 24 h *Rex1*^{low} and the 9 cells of 48 h. Chromosome structures were assumed to be ellipsoids in the analysis, to simplify the geometric realisation of their *I* ratios.

Figure 5.2.1 shows the distributions of *I* ratios for each chromosome, gathered from all cells in each time point and condition. Similar to mESCs, cells at all three time points and conditions have varied *I* ratios for each chromosome, indicating varied chromosome shape in individual cells. The majority of *I* ratios are below 2.5, with a few outliers mainly found in 24 h *Rex1*^{high} and 24 h *Rex1*^{low} chromosomes. Chromosome 18 of 24 h *Rex1*^{low} Cell 4 has the highest *I* ratios of over 4, indicating the most extended pseudo-axis of the structure is more than four times longer than the other two axes.

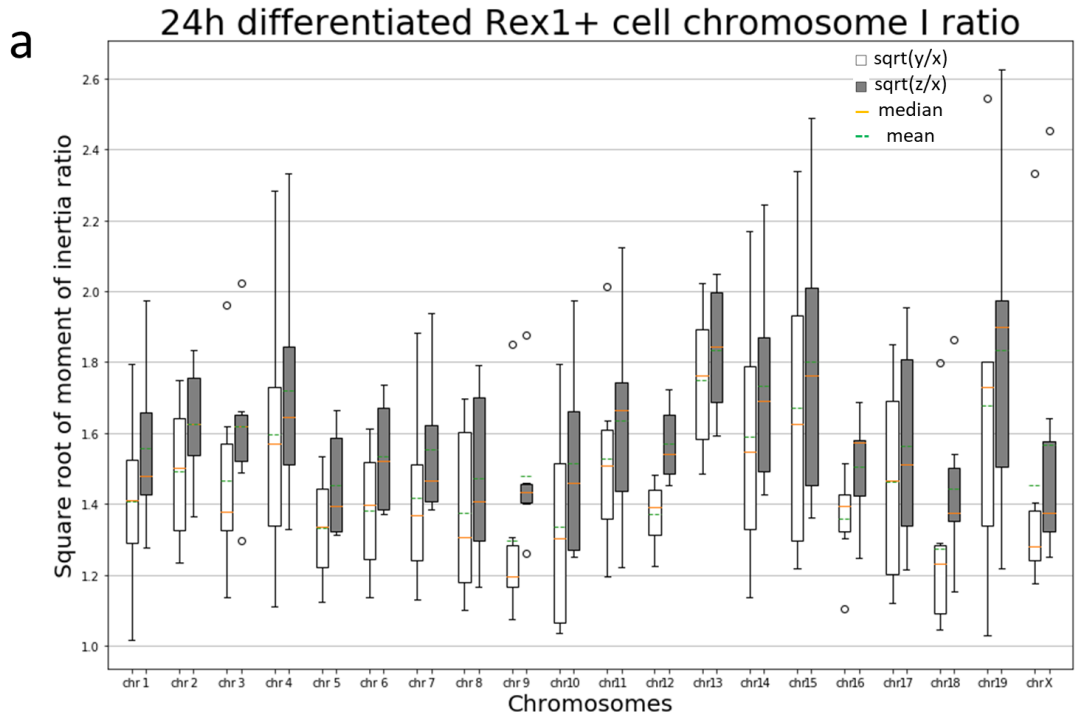


Figure 5.2.1 Chromosome *I* ratio distributions for differentiated cells

(See the next page for the rest of the figure and figure legends)

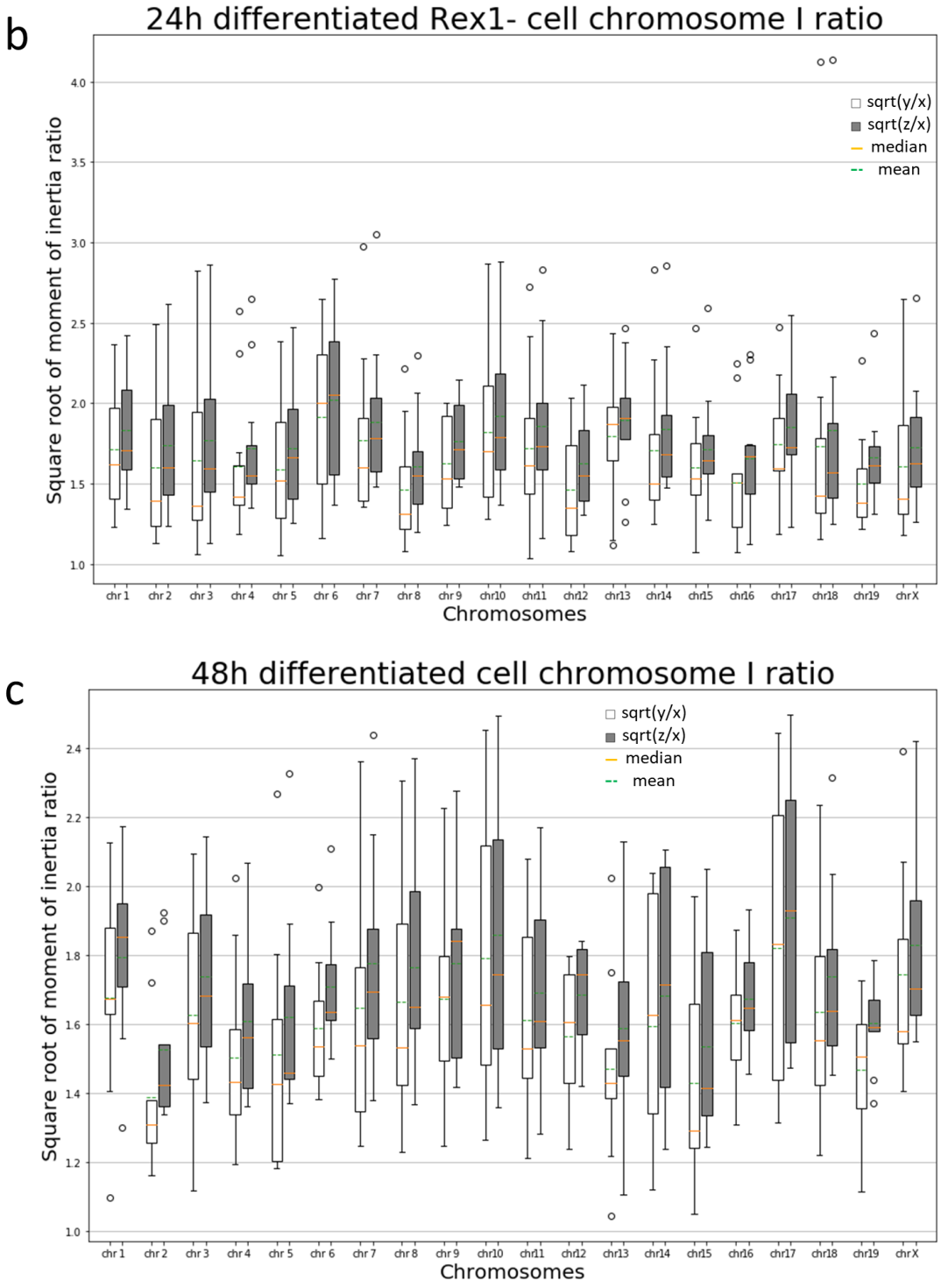


Figure 5.2.2 Chromosome *I* ratio distributions for differentiated cells

Box plot of each chromosome's *I* ratios of the 6 cells of 24 h *Rex1*^{high} (a), the 11 cells of 24 h *Rex1*^{low} (b) and the 9 cells of 48 h (c). White and grey boxes represent sqrt(I_y/I_x) and sqrt(I_z/I_x) respectively. Circles represent the outliers in the corresponding distributions. Within each box, the orange line represents the median and the dashed green line represents the mean.

Statistical t-test was used to compare the chromosome *I* ratios from the three time points and conditions and with mESC chromosomes. For better indications of structure flatness, sqrt(I_y/I_x) and sqrt(I_z/I_x)-sqrt(I_y/I_x) (the difference between the two *I* ratios) were analysed rather than sqrt(I_z/I_x). In order to choose the correct type of test, all distributions were first analysed for normality and all distribution pairs to be compared were analysed for equal variance. Due to low sample sizes, the normality was analysed using probability plots and their correlation coefficients (PPCC) for normal distribution at 1% significance level (Figure 5.2.2 a) [100,101]. Although some datasets did not pass the PPCC normal distribution test, because the sample sizes were low, it was still assumed that all datasets were normally distributed. On the other hand, typical statistical tests for equal variance are not accurate for dataset groups of small sample sizes. As a common recommendation for this case, the group variance would be assumed to be equal if the largest variance in the group is less than three times the smallest variance[102,103]. Based on the comparison shown in Figure 5.2.2 b, most of the chromosome *I* ratio dataset groups have the largest variance significantly exceed three times the smallest variance. Thus it was not feasible to assume the group variances to be equal. For normal distributed chromosome *I* ratio datasets with unequal variances, Welch's t-test was used instead of the typical Student's t-test. As a large number of dataset pairs were to be compared, and no significant trend can be seen from the distribution box plot (Figure 5.2.1), two-tailed test was used for all these comparisons.

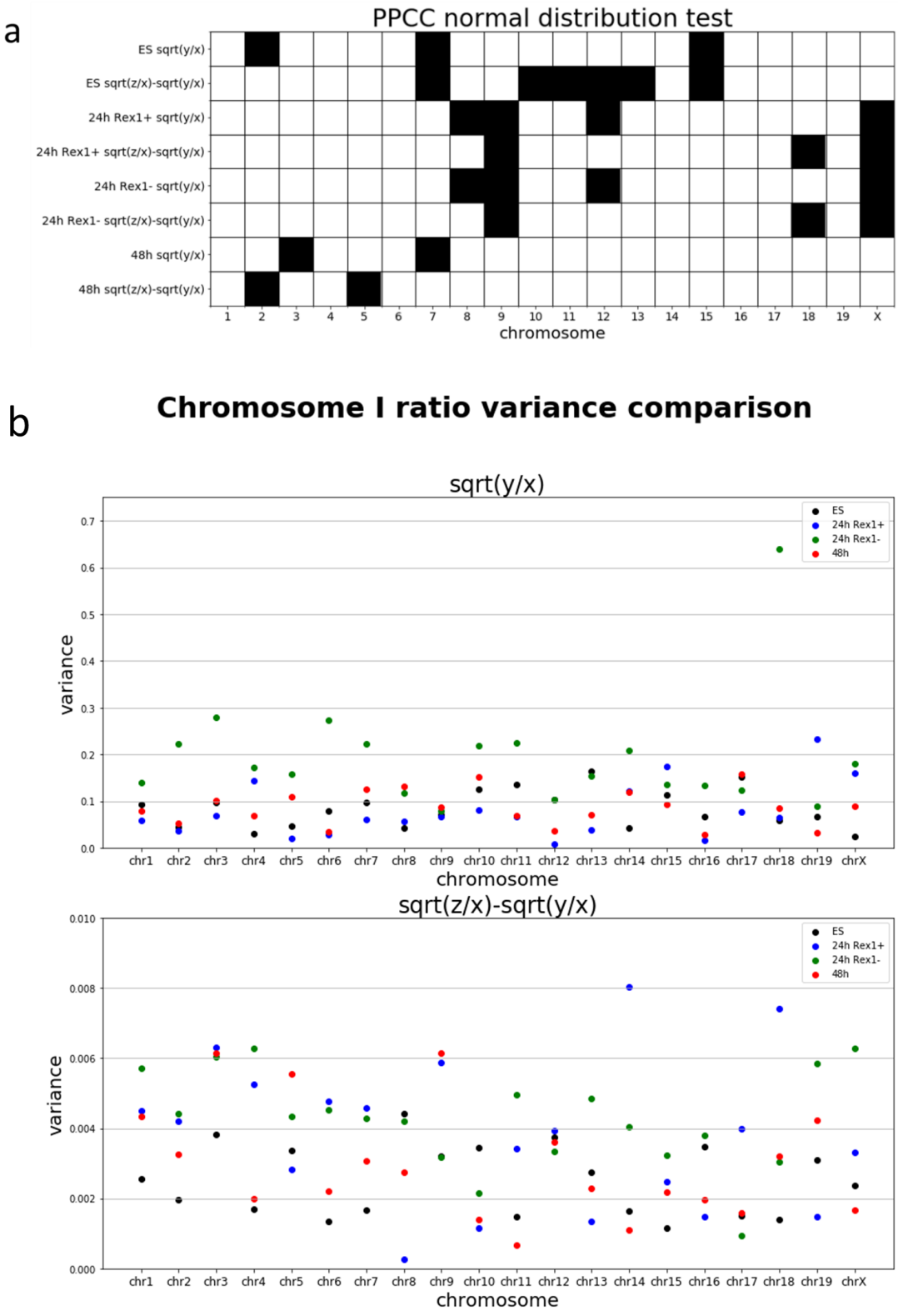


Figure 5.2.2 Assumption checks for t-test on chromosome *I* ratios

a) Normality analysis on chromosome *I* ratio distributions by PPCC normal distribution test (lower one-tailed, 1% significance level). Each block represents the result of the analysis on the corresponding distribution, where white indicates a pass and black indicates a fail. b)

Scatter plots of chromosome *I* ratio (sqrt(I_y/I_x) on the top and sqrt(I_z/I_x))-sqrt(I_y/I_x) at the bottom) dataset variances. Each chromosome holds a group of datasets to be compared. Each dataset is a spot on the plot, coloured by its time point and condition.

The results of the Welch's t-test are summarised in Figure 5.2.3. Most chromosomes have equal *I* ratios (sqrt(I_y/I_x) and sqrt(I_z/I_x)-sqrt(I_y/I_x)) for all four time points and conditions. The only differences in chromosome sqrt(I_y/I_x) values occur between 48 h and other time points and conditions, where 24 h *Rex1*^{high} and 24 h *Rex1*^{low} are more different from 48 h than ES. Within these comparisons, sqrt(I_y/I_x) values of chromosomes 9, 10, 12, 16 and 18 are the most different. This indicates that these chromosomes were stretched on one of the three dimensions to a different extent when the cells were differentiated for 48 hours. Interestingly, the differences in chromosome sqrt(I_z/I_x)-sqrt(I_y/I_x) values also occur in chromosomes 10, 12 and 16, which are mainly found between both conditions at 24 h and ES, and between both conditions at 24 h and 48 h time point. This indicates that these chromosomes were flattened, or stretched on two of the three dimensions, to a different extent when the cells were differentiated for 24 hours.

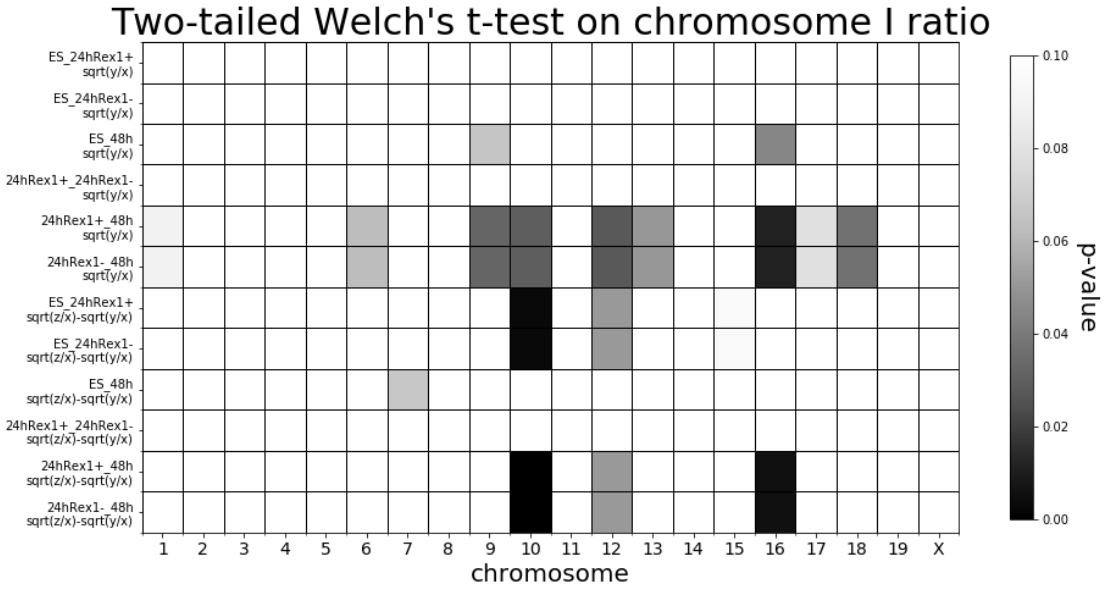


Figure 5.2.3 Results of two-tailed Welch's t-test on chromosome *I* ratios

Each block represents the p-value of the test on the corresponding comparison, where white

indicates a p-value of over 0.1 and grey-black indicates a p-value between 0 and 0.1 as shown in the colour bar.

5.3. Distinct genome flatness at different time points and conditions

Similar to chromosomes, genome flatness of cells from at all time points and conditions was also analysed using the moment of inertia (*I*) strategy described in Section 4.4. This analysis used the same cells involved in the chromosome flatness analysis as discussed in Section 5.2. The three *I* values were averaged from 10 structure models of each cell, before being used to calculate *I* ratios ((sqrt(I_y/I_x) and sqrt(I_z/I_x)). To simplify geometric realisation, all genome structures were assumed to be ellipsoids during the analysis.

As illustrated in Figure 5.3.1, each time point and condition shows a unique pattern of genome *I* ratio distribution. In general, ES, 24 h *Rex1*^{low} and 48 h have similar sqrt(I_y/I_x) distributions, which are significantly higher than that of 24 h *Rex1*^{high}. However, 24 h *Rex1*^{high} has the largest gap between the two *I* ratio distributions, which results in one of the highest distributions of sqrt(I_z/I_x), comparable to that of 24 h *Rex1*^{low}. It should note that the apparently most confined distribution of 24 h *Rex1*^{high} can be due to its smallest sample size (6 compared to 11, 11 and 9).

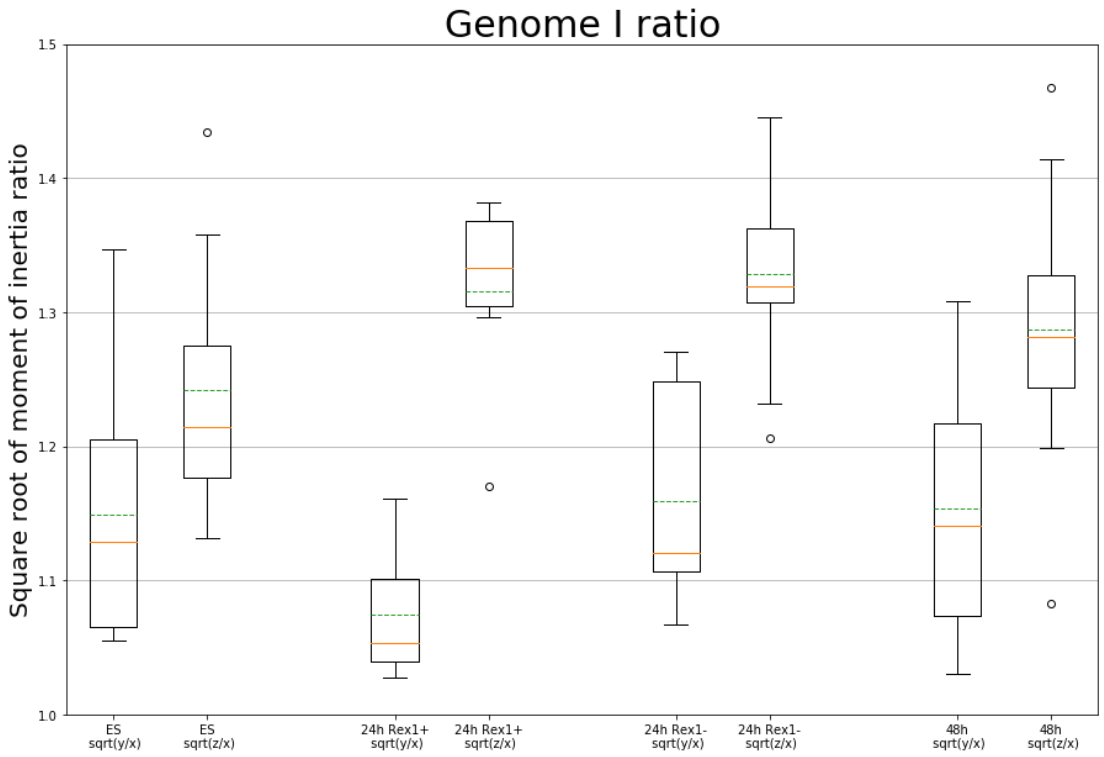


Figure 5.3.1 Genome I ratio distribution for all four time points and conditions

Box plot of genome *I* ratios of the 11 ES cells, the 6 24 h *Rex1*^{high} cells, the 11 24 h *Rex1*^{low} cells and the 9 48 h cells. White and grey boxes represent sqrt(I_y/I_x) and sqrt(I_z/I_x) respectively. Circles represent the outliers in the corresponding distributions. Within each box, the orange line represents the median and the dashed green line represents the mean. Data shown in Figure 4.4.3 are also included in this figure.

Statistical t-test was also used to compare genome *I* ratios as already shown for chromosome *I* ratios in Section 5.2. Again, distributions of sqrt(I_z/I_x)-sqrt(I_y/I_x) (the difference) were analysed along with sqrt(I_y/I_x) instead of sqrt(I_z/I_x). Normality and equal variance for these distributions were also analysed before the actual t-test, using the same methods as described in Section 5.2. Because 6 out of 8 datasets passed the PPCC normality test and the sample sizes were low, it was assumed that all datasets were normally distributed (Figure 5.3.2 a). Different from chromosome *I* ratio datasets, the ratio of the largest variance to the smallest variance for sqrt(I_y/I_x) is less than 3 and that for sqrt(I_z/I_x)-sqrt(I_y/I_x) is less than 5. The relatively larger ratio of sqrt(I_z/I_x)-sqrt(I_y/I_x) variance is partially due to the smallest variance of 24 h *Rex1*^{high},

which has the smallest sample size. Thus it was assumed that the datasets have equal variance. For normal distributed datasets with unequal variances, the typical Student's t-test was used for genome *I* ratios. The datasets were then sorted by descending order of their mean, and one-tailed test was used to analyse whether a former dataset is larger than a later one.

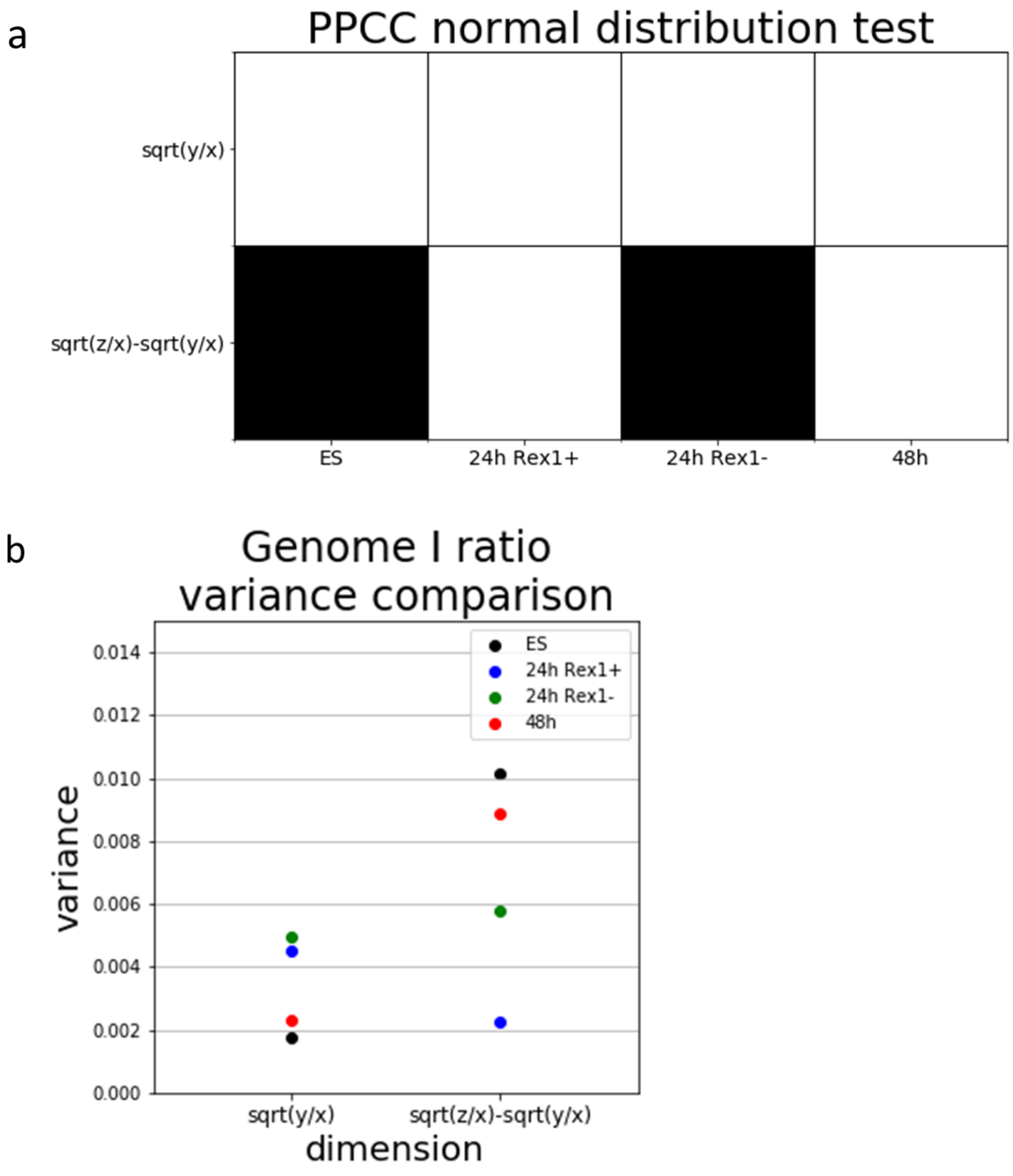


Figure 5.3.2 Assumption checks for t-test on genome *I* ratios

a) Normality analysis on genome *I* ratio distributions by PPCC normal distribution test (lower one-tailed, 1% significance level). Each block represents the result of the analysis on the

corresponding distribution, where white indicates a pass and black indicates a fail. b) Scatter plots of genome *I* ratio (sqrt(I_y/I_x) and sqrt(I_z/I_x))-sqrt(I_y/I_x)) dataset variances. Each *I* ratio holds a group of datasets to be compared. Each dataset is a spot on the plot, coloured by its time point and condition.

Figure 5.3.3 summarises the Student's t-test results. As shown in the top map, 24 h *Rex1*^{low}, 48 h and ES all have similar sqrt(I_y/I_x) values. However, values of 24 h *Rex1*^{high} are significantly lower than 24 h *Rex1*^{low} values (p-value 0.017) and moderately lower than those of 48 h and ES (p-values 0.0507 and 0.0643 respectively). These indicate that 24 h *Rex1*^{high} genome structures are more stretched in one of the three dimensions compared with the other time points and conditions. More interestingly, the bottom map shows that the sqrt(I_z/I_x)-sqrt(I_y/I_x) values significantly vary between most of the compared time point and condition pairs. Following the order of 24 h *Rex1*^{high}, 24 h *Rex1*^{low}, 48 h and ES, the values drop significantly at each step (p-values all <0.05) except from 24 h *Rex1*^{low} to 48 h, which still shows a slight decrease (p-value 0.1162). These indicate that the genome structures are the most flattened at 24 hour differentiated, *Rex1*^{high} state, and decreasingly flattened at 24 h *Rex1*^{low}, 48 h and ES states. All these observations were verified in the calculated structures, as exemplified in Figure 5.3.4.

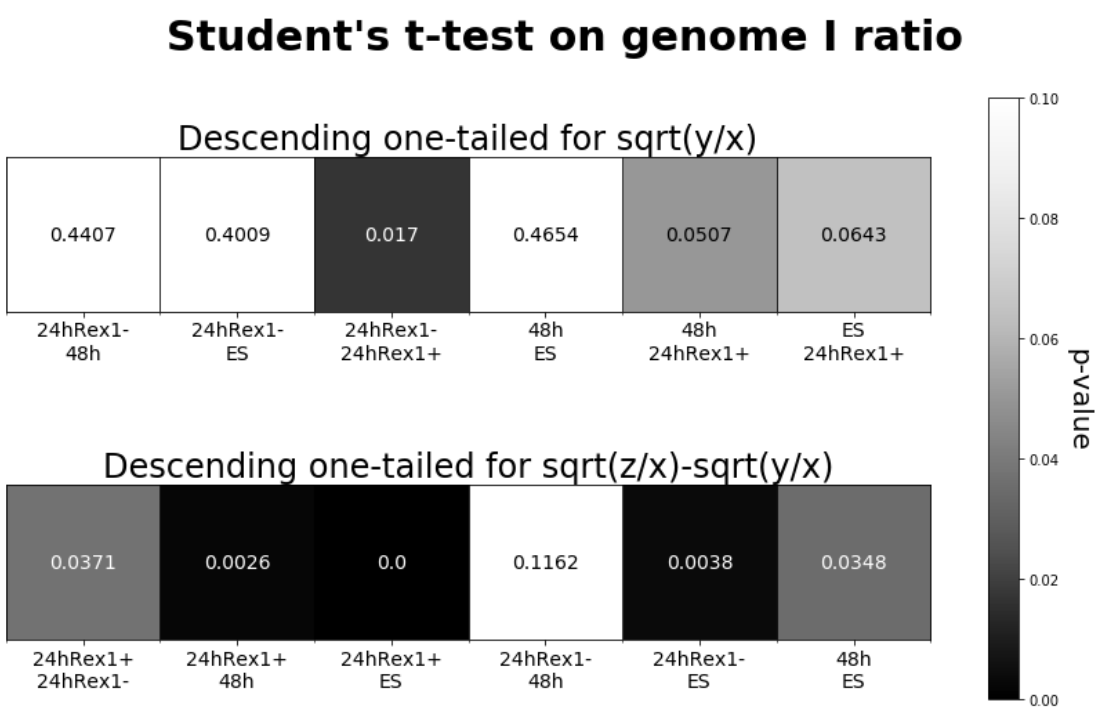


Figure 5.3.3 Results of one-tailed Student's t-test on genome *I* ratios

For each one-tailed comparison, the dataset with larger mean is labelled over the dataset with smaller mean. Each block represents the p-value of the test on the corresponding comparison, where white indicates a p-value of over 0.1 and grey-black indicates a p-value between 0 and 0.1 as shown in the colour bar. The actual p-values are shown in the corresponding block, where values smaller than 0.05 are coloured in white and values equal to or greater than 0.05 are coloured in black.

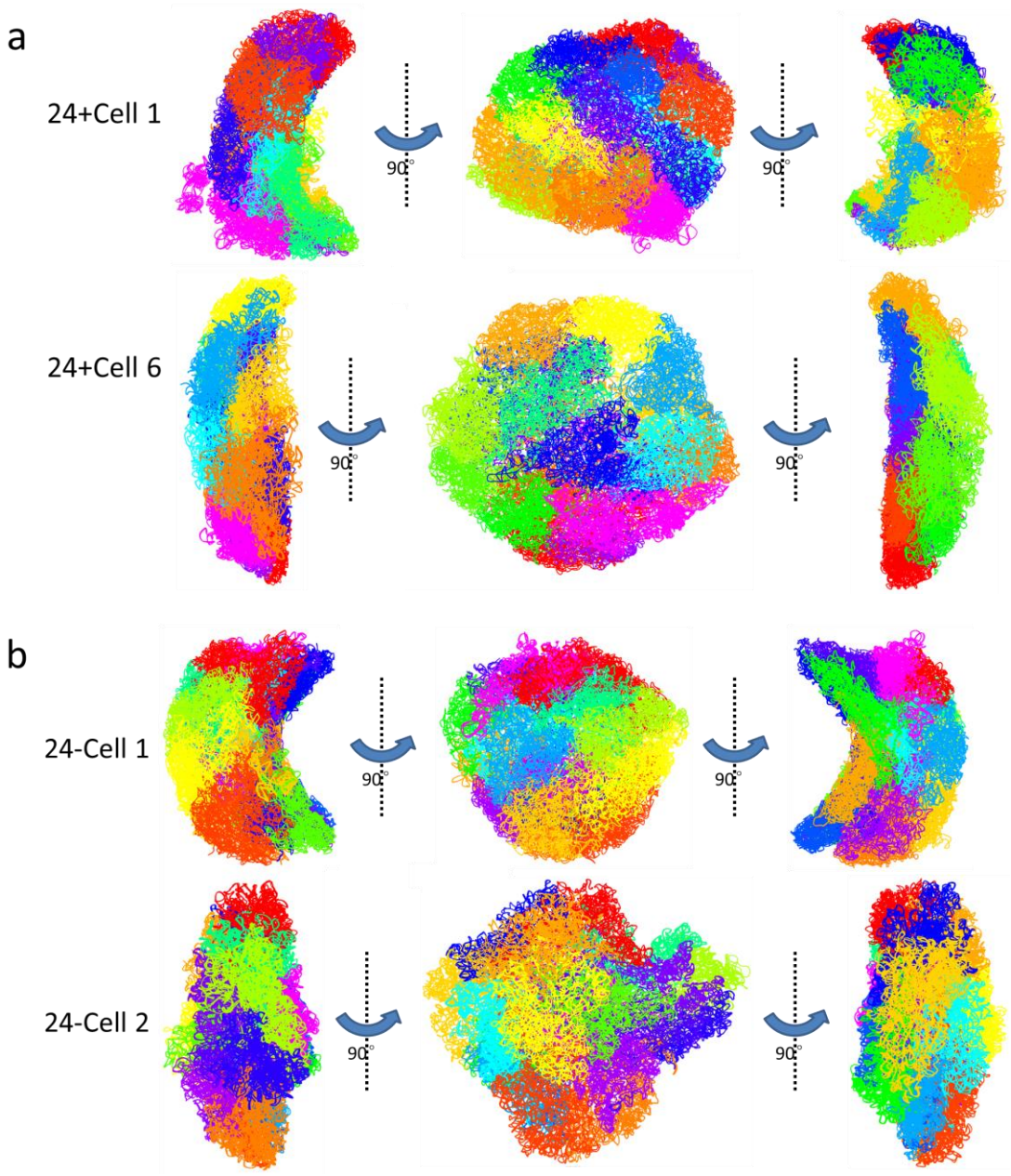


Figure 5.3.4 Genome structure examples of differentiated cells.

(See the next page for the rest of the figure and figure legends)

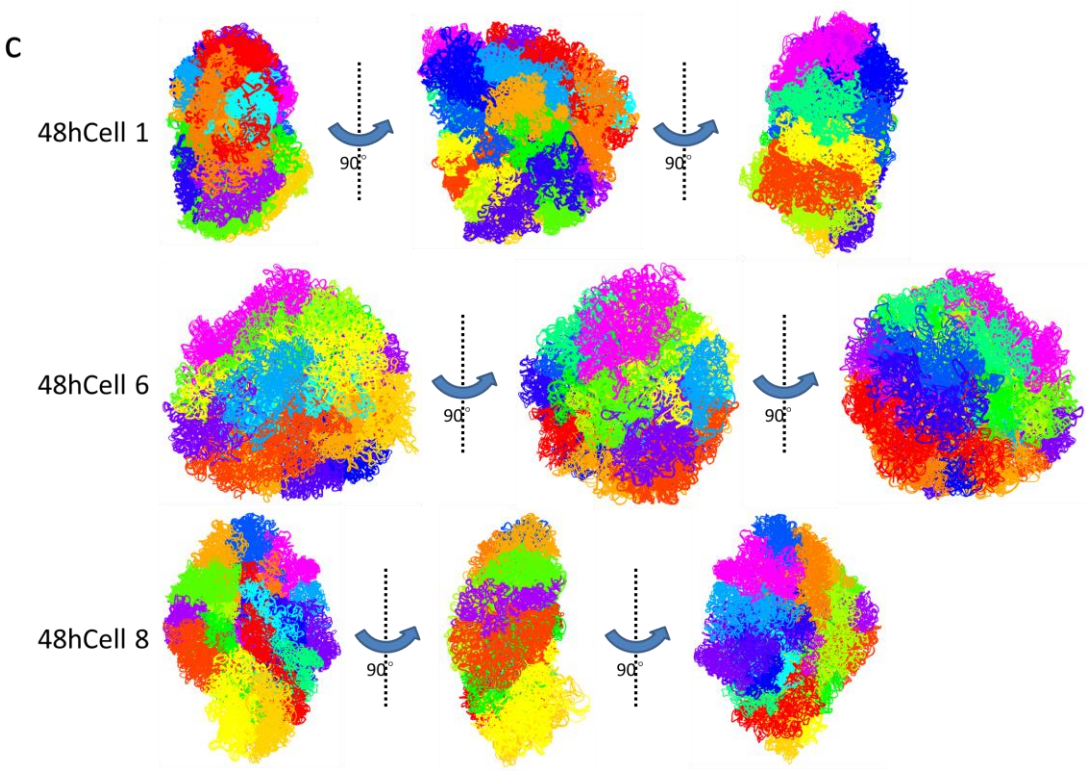


Figure 5.3.4 Genome structure examples of differentiated cells.

Calculated whole genome structures of 24 h *Rex1*^{high} Cell 1 and 6 (a), 24 h *Rex1*^{low} Cell 1 and 2 (b) and 48 h Cell 1, 6 and 8 (c). Each structure contains 5 overlaid models and coloured by chromosomes.

6. Further discussion and future work

6.1. Transposase method development

As discussed in Chapter 3, I aimed to adapt the transposase system for sequencing library preparation to our combined imaging single-cell Hi-C protocol. To achieve this goal, I made several critical modifications to the protocol of the commercial Nextera XT transposase kit. The modified protocol was used to successfully process two single-cell Hi-C libraries, but was not capable of giving consistently results. However before the method development had to pause due to time limitations, I was able to consistently process 2.5 pg (single nucleus equivalent) population Hi-C control samples to give libraries with good sequence quality. Here I suggest several potential causes of the inconsistency in single-cell samples and propose possible solutions.

6.1.1. The accessibility hypothesis

The accessibility of DNA to the transposase for tagmentation drew increasing attention throughout this method development project. There is no direct experimental evidence as of yet, so it is currently still a hypothesis, but here I summarize and discuss the relevant findings.

As discussed in Section 3.3.1, when liquid agarose is present during tagmentation of purified population Hi-C DNA, it moderately inhibits transposase activity leading to extended fragments. I suggest that this result can be explained by the viscous agarose solution slowing transposase diffusion and targeting. This can also explain the finding that higher agarose concentrations had a slightly stronger inhibiting effect. The transposase is thought to bind, cut and then ligate for a robust one-shot mechanism. In the five-minute tagmentation, the slow targeting would explain the results. This experiment suggests that robust transposase binding and tagmentation requires easy and immediate association with target DNA. It could also be that it inhibits the

interaction of the transposase with DNA.

Because tagmentation does not fragment the DNA[84,93], unlike AluI digestion, it has a limited effect on opening up chromatin structure. To achieve good coverage over the whole sample, ideally all possible sites of the DNA should therefore be available for the transposase to directly target. This requires that input DNA is all in a fully open conformation. The limited coverage observed suggests that the DNA was not in an open conformation, and that the buried interior was not accessible to the transposase. The reason for this is thought to be that the previous reactions (fixation, restriction enzyme 1 digestion, biotin end-filling and junction ligation) all kept the DNA around the original nucleus region for higher reaction efficiency, until the crosslink removal step which had limited effect on opening the conformation. So for single-cell Hi-C samples, particularly when processed with an agarose pad, it might be necessary to open the chromatin structure for better tagmentation efficiency.

As discussed in Section 3.4.1, over-tagmentation was thought to involve repeated tagmentation of adaptors at a preferred distance from the biotinylated Hi-C junction, and as shown in Section 3.4.2 over-tagmentation cannot be removed by solely reducing the amount of transposase. One possible explanation could be that untrimmed single-nucleus DNA only exposed regions around a few biotinylated Hi-C junctions for the transposase to target. The rest of the DNA including other biotinylated Hi-C junctions was somehow buried in the chromatin structure and inaccessible, so the reduced amount of transposase was still in excess. A possible mechanism was suggested in Section 3.4.1 and Figure 3.4.1.1 c. This conclusion is consistent with evidence that trimming facilitated access to on-bead single-nucleus DNA and avoided over-tagmentation thereby improved tagmentation coverage. This can possibly be explained by the AluI, a true restriction enzyme, opening up chromatin structure by cutting DNA into successively smaller pieces.

Tagmentation was not the only reaction in Hi-C experiments with a concern in DNA accessibility. Sequence analysis of some population Hi-C samples processed at the same time as the single-cell samples suggested a significant defect generating valid Hi-C junctions. This could potentially be due to inefficient reactions by the restriction

enzyme (referred as restriction enzyme 1 or RE1, MboI in our case) that creates the sticky ends, by the Klenow enzyme for biotin end-filling, or by the ligase. Fewer Hi-C junctions generated in the first place would directly reduce the sequence variation. Similar to tagmentation, the inefficiency in RE1 digestion is thought to be correlated with DNA inaccessibility. So it is possible that the DNA inaccessibility and inefficient RE1 digestion problems were also present in the single-cell Hi-C experiments.

The RE1 digestion inefficiency in our population Hi-C experiments had recently been improved and consistent results have been achieved by using different fixation protocols for cells at different stages of differentiation. These protocols with improved consistency are also benefiting our single-cell Hi-C experiments using the AluI-A-tailing method. Future work on the transposase method development can also rely on these protocols, which could help with investigating the true source of inconsistency and making further improvements.

6.1.2. Further discussion on input DNA with agarose

Apart from the inhibition effect on tagmentation, agarose also made it difficult to control the volumes for the transposase reactions. In single-cell Hi-C experiments, the minimum volume of the agarose pad is 20 µL. This volume ensures that the bottom of the well and nuclei are entirely covered by the pad. During the various Hi-C processing steps it is often also hard to completely remove all the residual solution from previous reactions when an agarose pad is present. Hence if the agarose is not removed before tagmentation, the input volume would be at least 20 µL, with potential variations if residual solution remains. It is also not possible to measure the input volume before tagmentation using any available apparatus in the current experimental setup. Therefore the current single cell Hi-C protocol is incompatible with the standard Nextera protocol where a 5 µL sample and final reaction volume of 20 µL is required. This is a particular problem because the concentration of transposase in the tagmentation reaction is likely critical for generating the correct

fragment size distribution in single-cell Hi-C experiments (as discussed in Section 3.3.3).

As discussed in Section 3.2 and above, agarose is incompatible with tagmentation and thus the transposase single-cell Hi-C method. My results suggested that it was necessary to remove agarose from Hi-C DNA before tagmentation. This could only feasibly be done by binding biotinylated Hi-C DNA to streptavidin-coated magnetic beads and washing away the agarose and exchanging the buffer. As a consequence, tagmentation would need to be carried out with Hi-C DNA bound on the bead, where the effects of magnetic beads on the transposase method should be tested (see Section 3.3 for corresponding results). However, in the traditional AluI-A-tailing method this bead purification step was finished before AluI restriction digestion (see Section 2.2 for a detailed comparison of the two methods). This suggested that agarose in input DNA could be removed by means of the magnetic beads so that it did not affect the following reactions. It should also be pointed out that it is important to effectively eliminate agarose because any remaining would solidify and probably stay in the sample throughout this stage until the PCR-amplified library purification. According to my experiences, it would not only stiffen the sample and make the handling more difficult, but also affect the library quality (data not shown). This was true for both the transposase reaction and the traditional AluI-A-tailing reactions.

6.1.3. Further thoughts on trimming

As discussed in Section 3.4.2, trimming allows effective tagmentation and reduces the uninformative fragments in the library. The current choice of trimming is AluI restriction, which is also used to fragment DNA in the AluI-A-tailing method. AluI restriction is suitable for the AluI-A-tailing method, because it mostly generates DNA fragments in the optimal size range 300 – 700 bp. (This includes the base pairs between the two AluI restriction sites, plus the two adaptors ligated to them.) However, in the case of tagmentation after trimming, the transposase randomly targets

DNA and adds adaptors between the cleavage sites. This further trims target fragments, which may leave them with too little sequence information such that may map to multiple sites of the genome or may even be unmappable. This problem may be compensated by a restriction enzyme with a less frequent recognition site than AluI site, giving generally longer fragments for tagmentation. However, the extended sequence may not have a biotinylated Hi-C junction and may be tagmented on its own. The transposase does not fragment DNA so these non-biotinylated fragments could not be removed during biotin purification. Thus this would increase the risk of generating more uninformative fragments in the library. The unfragmented DNA may be linked by either the unremoved transposase complex or a shared adapter sequence, or both. In future experiments, the transposase complex could potentially be removed by phenol or SDS heat treatment as suggested by Goryshin et al.[76]. To cleave the shared adapter sequence, home-made transposase with modified adapter sequences that contains an additional restriction site may be used with the corresponding restriction endonuclease as suggested by Parkinson et al.[88]. But the effects of these additional reactions on the tagmented single-cell DNA should be carefully tested. Other methods of trimming such as sonication could also be considered. This could avoid the potential bias caused by restriction enzymes with specific recognition sites, but it may be more difficult to obtain a relatively confined distribution of resultant fragment size. Due to time limits no comparisons between these methods have been made yet. Comparison with an unbiased method will be required at some stage to investigate any possible effects.

In addition, trimming and the following biotin purification reduces the amount of DNA in a single nucleus sample, which, however, is critical to adjusting the amount of transposase. This will be further discussed in the next section. In any case, trimming helps avoid over-tagmentation and achieve better sample coverage for tagmenting on-bead DNA. As shown in Section 3.4.1, streptavidin-coated magnetic beads tend to prevent the transposase from targeting sites that are too close to the biotinylated Hi-C junction. In the absence of over-tagmentation, the use of these beads with trimming may be beneficial avoiding too short fragments.

6.1.4. Further thoughts on the amount of transposase in tagmentation

Apart from trimming, the amount of the transposase is the other factor that affects tagmented fragment size and the amount of uninformative fragments. In theory, as they randomly target DNA, fewer transposases are more likely to generate longer biotinylated fragments and fewer uninformative fragments, but as discussed in Section 3.4.3, too low concentration does reduce tagmentation. Although the optimal relative concentration of Nextera transposase has been set to 1/100 of the suggested amount thus far, the ultimate dilution requires further consideration of a few factors.

As discussed in 3.4.2 and 6.1.3, trimming the DNA and the following biotin purification, further reduce the amount of DNA for tagmentation to an unknown amount. If the cell culture and sorting successfully prepares single haploid G_1-phase mouse embryonic stem cell (mESC) samples, the amount of DNA would also depend on the efficiencies of the Hi-C reactions and the biotin purification. So the amount of input DNA in tagmentation is extremely hard to control because all these factors can vary. It is also not viable to quantify the amount of on-bead DNA in the order of picogram. However, given that too diluted transposase will be less active, and the minimal amount (1/100 dilution) is still in excess compared with a single-genome amount of DNA, it should provide enough activity regardless of how much DNA is left.

Another possible source of transposase method inconsistency is the variation of transposase activity in different experiments. For each single-cell Hi-C sample I used only minimal amount of transposase for tagmentation, and the non-kit-based KAPA HiFi PCR polymerase for library amplification. Also the Nextera XT kit is rather expensive, so one kit was used for many experiments during several months. The activity of transposase might be decaying during this period; and in different periods transposase activity might have slight variations among different batches of the

Nextera XT kit. These potentially have significant effects on tagmentation in the single-cell Hi-C experiments. For example, if the transposase is 80% active compared with the first use, in theory one should use 25% more transposase to compensate for this activity defect, i.e. 1/80 of the suggested concentration instead of 1/100. This may not be an issue for massive tagmentation using the Nextera protocol which uses the original amount of transposase to tagment 1 ng of DNA[90], but both the concentration and activity of transposase are very sensitive in the small-scale tagmentation of single-cell Hi-C. Less active transposase of the same amount may lead to significant defects in tagmentation coverage. However, no direct evidence has been found to indicate a clear connection between transposase activity and library quality, probably because single-cell Hi-C experiments always have other sources of variation.

In general, the amount of transposase used in tagmentation aims for a balance between no over-tagmentation and good activity/coverage. It would be good to have a quantitative way of controlling transposase activity for future experiments. Instead of using the Nextera transposase, quality control to the home-made transposase may be simpler as it is easier and cheaper to produce in relatively large amounts. Also, instead of using the diluted Nextera kit transposase which neglects concentration changes of other unknown reagents in ATM (Amplicon Tagment Mix, Nextera XT kit), it might be better to dilute home-made transposase and control the amounts of other reagents. Importantly, the group who developed the Tn5 transposase production method also achieved successful tagmentation on picogram amounts of DNA, which thus shows great potential for our single-cell experiments[89].

6.1.5. Sample variation

Although the whole protocol, from cell culture to sequencing data analysis, has become much more consistent than during the early stages of the project, there were still significant variations between samples in the same experiment, and between different experiments. For example, only two successful libraries were obtained from

six samples processed in the same way in the same experiment. In addition, this experiment was the only one that provided at least a few good libraries in nearly 30 single-cell experiments, all processed using the same transposase protocol except minor optimisations.

In comparison, recent single-cell Hi-C experiments carried out using the improved AluI-A-tailing method were more consistent allowing us to make proper statistical conclusions. If the Fluorescence-activated cell sorting (FACS) is successful, a good experiment would provide about 20 good cells out of 40 that could be identified by imaging. After processing, about 10 cells in average would give libraries with fragment distributions that were promising for sequencing. And finally after sequencing, around 5 libraries would have enough useful Hi-C contacts to model a structure. Not such good experiments might have even lower success rates at any step of this process, and it was common to obtain less than 10 structures in total from 5 consecutive experiments. Such variation in recent single-cell Hi-C experiments was very likely to be in the past transposase experiments as well, and may be even more significant because of the relatively unimproved sample preparation protocol.

6.1.6. The current stage of the transposase method development is not far from success.

In general, it is hard to control experimental variabilities in single-cell projects. The effect of any slight change in input samples or reagents is likely to cause inconsistency, and improving the technique is often a good way to reduce result inconsistency. It is also easier to troubleshoot sources of inconsistency in some steps by making other parts of the protocol more reliable. However this requires systematic controls over the whole experiment and a relatively long time especially for a complicated protocol like single-cell Hi-C.

Given that now my group have a more reliable protocol for single-cell sample preparation and Hi-C reactions, it is my expectation that it would now be more

feasible to find the exact causes of inconsistency in the transposase method for single-cell Hi-C. Our experience successfully generated a good library using this method will be important for future works. The transposase method ought to be more efficient and lead to a higher number of contacts. Work from other groups has used transposase for similar projects so it may also help our own method development (see reference [67,88,89]). Overall, I think not many further optimisations are required to solve the remaining issues.

6.2. Single-cell Hi-C

It has only been five years since the first single-cell Hi-C publication[56]. Although the continuously developing method is still far from mature[66,104], it has proven to be a powerful tool to study genome organisation including cell-to-cell variation[67,92,105]. Here I discuss the current state of the single-cell Hi-C method and suggest some potential improvements. Further thoughts on some of our results about mESC chromatin structure (see Chapter 4) are also discussed, with suggestions on future work.

6.2.1. Experiment success rate

As discussed in Section 6.1.5, due to significant sample variation, a good single-cell Hi-C experiment starting with 40 sorted cells would result in around 5 libraries with enough Hi-C contacts for structure calculation. This success rate largely depends on cell sample quality, and for our experiments using haploid G_1-phase mESCs, sample quality is usually a significant issue due to uncertainties in haploid sorting. Cells entering S-phase result in ambiguous Hi-C contacts that cannot be used in structure calculations or data analysis. So the number of useful datasets that could be obtained from a haploid cell experiment largely depends on the ratio of G_1-phase cells in the sorted population. We routinely sorted for haploid cells during cell sample preparation,

but with the fast cell cycle it is possible for them to enter S-phase prior to fixation. We also used CENP-A imaging to identify obviously diploid cells by their number of centromeres, but this cannot distinguish early S-phase cells, whose ploidy is close to 1N as haploid G_1 cells. Currently, early S-phase cells can only be identified until their sequencing reads are processed at the very end of the experiment, where Hi-C contact maps demonstrate that chromosomes contact with too many other chromosomes, indicating that more than one copy of such chromosomes were present in the genome. Cells that are just entering S-phase can also be recognized through regions of some chromosomes having a higher than expected number of contacts. In addition, haploid G_1 cells in the same sorted population tend to enter S-phase together, and that almost all libraries in that experiment cannot be used for structure calculation. In general, as all cells in an experiment share the same procedure which is thought to be relatively consistent, their libraries tend to be of similar quality. In other words, a good experiment would result in four or five single-cell Hi-C structures, whereas a problematic experiment would usually have none. Also an experiment resulting in more structures is likely to give better structures. So in general the best structures tend to come from only a few experiments out of many.

6.2.2. Production capacity

Each single-cell Hi-C sample requires a full set of experimental processing and computational sequence analysis comparable to a population Hi-C experiment. The capability for processing single-cell Hi-C samples in parallel depends on available facilities, but typically caps at about 60 samples per experiment due to time and technical limits[66,104]. Given that each experiment normally takes more than a month from cell culture to sequencing and read processing, in theory, collecting 20 valid single-cell Hi-C datasets using the current procedure would take around 6 months. Obviously, more samples would be better for demonstrating cell-to-cell similarity and variability, and carrying out proper statistical analysis to deduce the status of the

whole cell population.

6.2.3. Further thoughts on combined imaging single cell Hi-C experiments

A unique feature of our combined imaging single-cell Hi-C experiment is that we can carry out imaging and Hi-C on the same single cell. This allows both independent quality controls on each cell and direct superpositioning of imaging data onto the corresponding remodelled genome structures. Currently, our group has mainly imaged fluorescently labelled CENP-A, a centromeric histone H3 variant, and histone H2B over the whole genome[66,92]. Firstly, we used both signals to verify the G_1 status of our isolated single cells. Signals of CENP-A should be bright dots and the number of dots in a single sample represents the number of centromere regions, or the number of chromosome copies. This helped us identify wells containing multiple cells, multiple copies of haploid G1-phase chromosomes, or lost chromosomes. Signals of H2B cover the whole genome and indicate its shape and size. This was used to identify cell that were damaged (centromeres outside the nucleus), or samples with multiple cells. In addition, we used the CENP-A signals to validate our modelled structures, by superimposing centromere positions determined from the two distinct types of data (see Section 4.2). Thus we can combine imaging and single-cell Hi-C data of the same cell. However, single-cell Hi-C provides a snapshot of genome conformation at the time when the cell is fixed, thus only the imaging data at that particular time point can be mapped to structure. This means that whilst it might be possible to image cells in vivo before fixation to obtain dynamic information, only the static image after fixation can be directly linked to single-cell Hi-C data.

Imaging of a cell population can also be used as a complementary method to single-cell Hi-C. For example, we found clusters of two key components of the nucleosome remodelling deacetylase (NuRD) complex, the CHD4 a chromatin-remodelling component, and MBD3 a component of the histone

deacetylase sub-complex, using super resolution imaging of fluorescently labelled (mEos3.2-tagged) proteins. This was used to verify the clustering of NuRD genes found by mapping ChIP-Seq data onto our single genome structures.

6.2.4. Further thoughts on population Hi-C and A/B compartments

A/B compartments are defined solely using population Hi-C data[17]. This level of genome structure reflects Hi-C contact frequencies in a cell population, and currently it has not been shown possible to deduce this from a single cell Hi-C dataset. The A/B compartments are correlated with a number of characteristics, including gene density, chromatin accessibility, transcriptional activity, histone marks, replication timing and relative locations in a nucleus (see Section 1.1.4 for details.) It is particularly interesting to investigate A/B compartments in individual cells from the same population, in order to help answer whether all these differences are consistent in single cells. Remarkably, we have shown that A/B compartments in single mESCs have a highly consistent conformation, where genomic regions belonging to the same compartment aggregate together and segregate from regions from the opposite compartment, forming a B-A-B alternating bowl shape at the whole genome level (see Section 4.5).

6.2.5. Further thoughts on TADs and CTCF/cohesin loops

As discussed in Section 4.6, in general, the previously defined TADs and CTCF/cohesin loops identified in population studies do not form in all of our single cells. However it should be emphasised that our single-cell Hi-C data captures only a snapshot of genome organisation at the time when the cells were just fixed. The method does not capture snapshots at other time points for the same cell, and thus the results do not include any dynamic information. The partial presence of TADs and CTCF/cohesin loops may indicate population variation where in some cells the

structures had not yet formed or never form during that particular G1 phase. Alternatively, it could also be possible that the genome structure is highly dynamic and loops/TADs are constantly forming and unfolding. Both explanations would be consistent with the loop-extrusion mechanisms, and it is possible that they are both true.

6.2.6. Further thoughts on other complementary methods

Our single-cell Hi-C data analysis has shown how to combine data from other methods such as cLADs, ChIP-Seq data to define histone modifications and transcription factor binding, and RNA-Seq data to study the relationship between genome structure and gene expression (see Section 4.7). However, it should be noted that these data were all obtained from a population of cells rather than the same cell used for single-cell Hi-C. Similar to the A/B compartment analysis, it was remarkable to find a consistent organisation of these data across all single cells studied. These include the association of cLAD with nuclear membrane and nucleolar periphery, the correlations and anti-correlations of certain types of histone modifications, the clusters of active enhancers and promoters, and the preferred location of active genes at chromosome interfaces. All these relatively local mechanisms may provide the driving force for the formation of whole genome architecture. In general, like the CENP-A data from single-cell imaging (Section 6.2.3), these complementary data would potentially be even more informative if they are derived from single-cell experiments ideally on the same cell for single-cell Hi-C. However this will need substantial method development in the future.

6.3. Future work on differentiated cells

6.3.1. Some methods require minor optimisation.

In the relatively short term, our single-cell Hi-C experiments will focus on early mESC differentiation. I am contributing to collecting data at each differentiation time point and condition including 0 h (24 h after LIF removal), after 24 h differentiation (24 h after 2i removal) with either low *Rex1*-GFP or high *Rex1*-GFP abundancy, and after 48 h differentiation. To allow an analysis similar to what we have carried out on mESCs, the types and amount of data at each time point and condition should be comparable to the data we collected for mESCs. These include more single-cell Hi-C datasets, imaging data of interesting proteins, population Hi-C data to define A/B compartments, ChIP-Seq of interested proteins and RNA-Seq data. As discussed in Section 5.1, we have successfully collected a reasonable number of single-cell datasets at the 24 h *Rex1*-low, 24 h *Rex1*-high and 48 h time points and conditions. These cells were processed using the same procedure as for mESCs. However, we found that the population Hi-C protocol used for mESCs was not compatible with the 24 h and 48 h differentiated cells, where only about 1 million contacts could be identified from the processed library compared with over 50 million from a mESC population library. This issue has been solved by using a slightly different method as described in Rao et al.[35]. We are also trying to collect data from other complementary methods, which may also require optimisation for differentiated cells.

6.3.2. Studying changes in genome structure during differentiation

After collecting more data on differentiated cells, it will be interesting to analyse their genome structures in similar ways to mESC analysis. Preliminary analysis has shown significant differences in the flatness of the whole genome between mESCs and differentiated cells (see Sections 4.4 and 5.3). However the small sample sizes at all

time points and conditions limit the tests' statistical power. Collecting more single cell data would help improve this problem. Comparison on other aspects of the genome structure will allow us to further investigate the changes in genome structure during cell differentiation and their relationship to changes in function. It is also possible to carry out single-cell Hi-C for other time points and conditions such as 72 h after differentiation and for terminally differentiated cells, to further understand the differentiation process. And by doing this, it may also be possible to start studying large-scale DNA dynamics resulting from the larger cell cycle or in post-mitotic cells.

7. Methods and materials

The current single-cell Hi-C method that combines microscopy imaging with Hi-C processing has about 100 individual steps including both wet-lab and computational processing. From my experience, the full protocol from start to finish takes at least 3 weeks, which does not include time for cell culture or massively parallel high-throughput sequencing. The majority of this protocol has been published in the methods section of Stevens et al.[92], with a more recent and detailed version in the protocol paper by Lando et al.[66]. I also carried out some population Hi-C experiments on the cell populations used for single-cell Hi-C for complementary information like the A/B compartments. The materials and procedures largely resemble the single-cell Hi-C protocol and are outlined in Section 7.3.

7.1. Materials

7.1.1. Reagents

- 2i growth medium (NDiff B27 base medium containing 1 μM PD0325901, 3 μM CHIR99021 and 20 ng/ml LIF)
- Accutase reagent (Gibco, cat. no. A1110501)
- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63881)
- AluI (10,000 U/ml; New England Biolabs, cat. no. R0137S)
- Biotin-14-dATP (0.4 mM; Invitrogen, cat. no. 19524016)
- BSA (20 mg/ml; New England Biolabs, cat. no. B9000S)
- BW buffer (2 × stock: 10 mM Tris-HCl (pH 8), 1 mM EDTA and 2 M NaCl)
- BWT buffer (0.05% (vol/vol) Tween 20 in 1 × BW buffer (5 mM Tris-HCl (pH 8), 0.5 mM EDTA, 1 M NaCl))
- CHIR99021 (2 mg; Cambridge Biosciences, cat. no. 1677-25)
- Concentrated AMPure XP beads (Concentration upon specification: 1 × stock was

equilibrated to RT, beads separated from solution on a magnetic stand, certain amount of solution was removed by pipetting, and beads were resuspended in the remaining solution)

- cOmplete protease inhibitor (EDTA-free; Roche, cat. no. 11873580001)
- CutSmart buffer (10×; New England Biolabs, cat. no. B7204S)
- dNTP (dATP, dCTP, dGTP and dTTP) set (100 mM; Invitrogen, cat. no. 10297018)
  - dCTP/dTTP/dGTP mix (10 mM each of dCTP, dGTP and dTTP)
  - dNTP mix (10 mM each of dATP, dCTP, dGTP and dTTP)
- Dynabeads M-280 streptavidin (10 mg/ml; Invitrogen, cat. no. 11205D)
  - Dynabeads M-280 streptavidin bead slurry (resuspend Dynabead M-280 Streptavidin in doubled volume of 2×BW)
- EDTA (Sigma-Aldrich, cat no. E5134)
- Ethanol (96% pure; Honeywell, cat. no. 32294)
- Flow-Check fluorospheres (10 μm; Beckman Coulter, cat. no. 6605359)
- Formaldehyde (16% (wt/vol); Pierce, cat. no. 28908)
- Geneticin™ Selective Antibiotic (G418 Sulfate) (50 mg/mL; ThermoFisher, cat. no. 10131035)
- Gelatin (Sigma-Aldrich, cat. no. G1890)
- Glycine (Sigma-Aldrich, cat. no. 410225)
- Haploid mouse embryonic stem (mES) cells[65] (Sigma-Aldrich, cat. no. 14040203)
- High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626)
- Immersion oil (refractive index n = 1.518 at 23℃, 30 ml; Olympus, cat. no. IMMOIL-F30CC)
- KAPA HiFi Polymerase Kit (Kapa Biosystems)
- DNA polymerase I, Large (Klenow) fragment (5,000 U/ml; New England Biolabs, cat. no. M0210L)
- Klenow Fragment 3’→5’ exo- (5,000 U/ml; New England Biolabs, cat. no. M0212L)

- Library amplification primers (Particularly synthesised for Hi-C AluI-A-tailing method using HPLC-purified-grade oligonucleotides, see Table 7.1.1.2 for sequences)
  - Library amplification primer mix (25 mM each of forward and reverse primers; for AluI-A-tailing method)
- Lipofectamine 2000 Transfection Reagent (ThermoFisher, cat. no. 11668027)
- Low-melting-point agarose (Sigma-Aldrich, cat. no. A9414)
- MboI (25,000 U/ml; New England Biolabs, cat. no. R0147M)
- Mouse LIF protein (expressed and purified in-house)
- NaCl (Sigma-Aldrich, cat. no. 71376)
- NDiff B27 base medium (Stem Cell Sciences, cat. no. SCS-SF-NB-02)
- NEBuffer 2 (10×; New England Biolabs, cat. no. B7002S)
- NEBuffer 3 (10×; New England Biolabs, cat. no. B7003S)
  - NEBuffer 3 (50 mM Tris-HCl, 100 mM NaCl, 10 mM MgCl2 and 1 mM DTT, pH 7.9; alternative to the commercial NEBuffer 3, New England Biolabs)
- Nextera XT Library Preparation Kit (24 samples; Illumina, cat. no. FC-131-1024)
- Nextera XT Index Kit (96 indexes, 384 samples; Illumina, cat. no. FC-131-1002)
- NP-40 (IGEPAL CA-630; Sigma-Aldrich, cat. no. 18896)
- Nuclei extraction buffer (10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% (vol/vol) NP-40 and protease inhibitors (Roche)
- Oligonucleotide adaptors for sequencing, containing 3-letter barcodes (Particularly synthesised for Hi-C AluI-A-tailing method using HPLC-purified-grade oligonucleotides, see Table 7.1.1.1 for sequences)
  - Oligonucleotide adaptor (for each barcode: 12 mM in 0.1 M Tris-HCl (pH 8.0), 0.5 M NaCl; forward and reverse oligonucleotides (12 mM each) were annealed by heating to 95℃ for 5 min and gradually cooling down to RT; for AluI-A-tailing method)
- PBS (pH 7.4; Gibco, cat. no. 10010023)
- PD0325901 (2 mg; Cambridge Biosciences, cat. no. SM26-10)

- Platinum Pfx DNA Polymerase Kit (Invitrogen, cat. no. 11708021)
- Proteinase K (800 U/ml, equivalent to 20 mg/ml; New England Biolabs, cat. no. P8107S)
- Puromycin Dihydrochloride (ThermoFisher, cat. no. A1113802)
- Rainbow calibration particles (eight peaks; BioLegend, cat. no. 422903)
- SDS (Sigma-Aldrich, cat. no. 436143)
- sodium acetate (VWR chemicals, cat. no. 27652)
- T4 DNA ligase (400,000 U/ml; New England Biolabs, cat. no. M0202L)
- T4 DNA ligase buffer (10×; New England Biolabs, cat. no. 10297018)
- tandem iRFP tagged histone H2B plasmid (Miyanari et al.[106]; Addgene)
- mEos3.2-tagged CENP-A plasmid (Palayret et al.[107])
- Tris base (Sigma-Aldrich, cat. no. T4661)
- Tris-HCl (pH 8.0) buffer (1 M stock: 1 M Tris pH calibrated to 8.0 using HCl)
- Triton X-100 (Sigma-Aldrich, cat. no. T8787)
- Tween 20 (Sigma-Aldrich, cat. no. P9416)
- Ultrapure water, type 1 grade (Advantage A10 system; Millipore)

Table 7.1.1.1 Oligonucleotide adaptors (for AluI-A-tailing method)

Barcode	F/R[ab]	Sequence[cd]
AAC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAC*T-3'
AAC	R	5'-pGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
AAG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAG*T-3'
AAG	R	5'-pCTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
AAT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAT*T-3'
AAT	R	5'-pATTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
ACA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTACA*T-3'
ACA	R	5'-pTGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
ACC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTACC*T-3'
ACC	R	5'-pGGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

ACG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTACG*T-3'
ACG	R	5'-pCGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
ACT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTACT*T-3'
ACT	R	5'-pAGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
AGA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGA*T-3'
AGA	R	5'-pTCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
AGC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGC*T-3'
AGC	R	5'-pGCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
AGG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGG*T-3'
AGG	R	5'-pCCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
AGT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGT*T-3'
AGT	R	5'-pACTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
ATA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTATA*T-3'
ATA	R	5'-pTATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
ATC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTATC*T-3'
ATC	R	5'-pGATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
ATG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTATG*T-3'
ATG	R	5'-pCATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
ATT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTATT*T-3'
ATT	R	5'-pAATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CAA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAA*T-3'
CAA	R	5'-pTTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CAC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAC*T-3'
CAC	R	5'-pGTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CAG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAG*T-3'
CAG	R	5'-pCTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CAT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAT*T-3'
CAT	R	5'-pATGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CCA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCA*T-3'

CCA	R	5'-pTGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CCG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCG*T-3'
CCG	R	5'-pCGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CCT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCT*T-3'
CCT	R	5'-pAGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CGA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGA*T-3'
CGA	R	5'-pTCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CGC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGC*T-3'
CGC	R	5'-pGCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CGG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGG*T-3'
CGG	R	5'-pCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CGT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGT*T-3'
CGT	R	5'-pACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CTA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTA*T-3'
CTA	R	5'-pTAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CTC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTC*T-3'
CTC	R	5'-pGAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CTG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTG*T-3'
CTG	R	5'-pCAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
CTT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTT*T-3'
CTT	R	5'-pAAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GAA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAA*T-3'
GAA	R	5'-pTTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GAC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAC*T-3'
GAC	R	5'-pGTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GAG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAG*T-3'
GAG	R	5'-pCTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GAT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAT*T-3'
GAT	R	5'-pATCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

GCA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCA*T-3'
GCA	R	5'-pTGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GCC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCC*T-3'
GCC	R	5'-pGGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GCG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCG*T-3'
GCG	R	5'-pCGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GCT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCT*T-3'
GCT	R	5'-pAGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GGA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGA*T-3'
GGA	R	5'-pTCCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GGC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGC*T-3'
GGC	R	5'-pGCCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GGT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGT*T-3'
GGT	R	5'-pACCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GTA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTA*T-3'
GTA	R	5'-pTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GTC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTC*T-3'
GTC	R	5'-pGACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GTG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTG*T-3'
GTG	R	5'-pCACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
GTT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTT*T-3'
GTT	R	5'-pAACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TAA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAA*T-3'
TAA	R	5'-pTTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TAC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAC*T-3'
TAC	R	5'-pGTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TAG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAG*T-3'
TAG	R	5'-pCTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TAT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAT*T-3'

TAT	R	5'-pATAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TCA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCA*T-3'
TCA	R	5'-pTGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TCC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCC*T-3'
TCC	R	5'-pGGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TCG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCG*T-3'
TCG	R	5'-pCGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TCT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCT*T-3'
TCT	R	5'-pAGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TGA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGA*T-3'
TGA	R	5'-pTCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TGC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGC*T-3'
TGC	R	5'-pGCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TGG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGG*T-3'
TGG	R	5'-pCCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TGT	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGT*T-3'
TGT	R	5'-pACAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TTA	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTA*T-3'
TTA	R	5'-pTAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TTC	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTC*T-3'
TTC	R	5'-pGAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
TTG	F	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTG*T-3'
TTG	R	5'-pCAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

[a] Forward and reverse respectively

[b] Each pair of oligonucleotides (F) and (R) with the same barcode should be annealed to prepare the adaptor.

[c] “p” at the 5’ end of the oligonucleotides (R) indicates 5' phosphate modification.

[d] “*” near the 3’ end of oligonucleotides (F) indicates 5'-3' phosphorothioate linkage.

Table 7.1.1.2 Library amplification primers (for AluI-A-tailing method)

F [a]	5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3′ [c]
R [b]	5’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T-3′ [c]

[a] Forward

[b] Reverse

[c] “*” near the 3’ end of oligonucleotides (F) indicates 5'-3' phosphorothioate linkage.

7.1.2. Equipment

- 0.2-micron Minisart filter (Sartorius, cat. no. 16534K)
- 0.2-ml Low-bind PCR tube (Corning Axygen, cat. no. PCR-02-L-C)
- 1.5- and 0.2-ml Tube magnetic separation stand (Promega, cat. no. Z5342)
- 1.5-ml DNA Lo-bind tube (Eppendorf, cat. no. 0030108051)
- 15-ml Centrifuge tube (Corning Falcon, cat. no. 352096)
- 2100 Bioanalyzer system (Agilent Technologies, cat. no. G2939AA)
- 384-well Cyclo-olefin polymer (COP) clear-bottom plate (Porvair Sciences, cat. no. 327001)
- 384-well Plate magnetic separation stand (Promega, cat. no. V8241)
- 50-ml Centrifuge tube (Corning Falcon, cat. no. 352070)
- ARKTIC Thermal Cycler (Type: 5020, ThermoFisher)
- Bench top minicentrifuge (Gilson, cat. no. F110733)
- Filter (50 micron; Sysmex, cat. no. 04-004-2327)
- Finnpipette Novus electronic single-channel multidispenser pipette (Thermo Fisher Scientific, cat. no. PIP1708)
- Heraeus Megafuge 16 centrifuge (Thermo Fisher Scientific, cat. no. 75004230)
- HiSeq sequencing system (Illumina)
- Incubator (Kuhner, cat. no. SMX1501)

- Low-bind filter tips (Corning Axygen, cat. nos. TF-20-L-R-S, TF-200-L-R-S, TF-300-L-R-S and TF-1,000-L-R-S)
- Mastercycler nexus flat PCR machine (Eppendorf, cat. no. 6335 000.011)
- Microwave (Bosch, cat. no. HMT72G450B)
- MiSeq sequencing system (Illumina)
- MoFlo cell sorter equipped with a Cyclone unit (Beckman Coulter) mounted on an air-dampened table (Newport, cat. no. RS 2000, S-2000)
- Nanodrop Spectrophotometer (model ND-1000, ThermoFisher)
- PCR adhesive plastic film (Bio-Rad, cat. no. MSB1001)
- Pipetman (Gilson, volumes. and cat. nos. 0.2 to 2µL F144801, 1 to 10µL F144802, 2 to 20µL F123600, 50 to 200µL F123601 and 200 to 1000µL F123602)
- Pipette controller (Corning Falcon, cat. no. 357469)
- Plate mixer (Thermomixer) with 384-well plate block/adapter (Eppendorf, cat. nos. 5382 000.015 and 5307 000.000)
- Platform rocker (Thermo Fisher Scientific, cat. no. M79735Q)
- Serological pipettes (Corning Falcon, cat. nos. 357529 and 357530)
- Sonic Dismembrator with tips (model FB505, Fisher)
- T75 flask (Corning Falcon, cat. no. 353110)

7.1.3. Microscope

- Aperture (Thorlabs, cat. no. IDA20-P5)
- Computer (3.4-GHz, Intel Core i7, 16-GB RAM, 64-bit operating system; Dell, model no. OptiPlex 990, with Windows 7 Professional)
- Concave and convex lens to expand the beam (Thorlabs, cat. nos. LC1054 and LA1484)
- Dichroic beam splitter (Semrock, cat. no. Di01-R405/488/561/635)
- Dichroic mirror (Semrock, cat. no. FF458-Di02-25×36)

- EMCCD (electron-multiplying charge-coupled device) camera (Photometrics, model no. Evolve 512)
- Emission filters (488-nm long-pass, Semrock, cat. no. BLP01-488R; 520/35-nm band-pass, Semrock, cat. no. FF01-520/35; 641-nm long-pass, Semrock, cat. no. BLP01-635R)
- Excitation lasers (488 nm, 100 mW; Toptica, model no. iBeam Smart 488; 641 nm, 100 mW; Coherent, model no. CUBE 640-100C)
- Image lens (×2.5 achromatic; Olympus, PE 2.5×125)
- Inverted microscope (Olympus, model no. IX71)
- Mechanical shutters for pulsing lasers (Prior, cat. no. HF202HT)
- Mirror (Thorlabs, cat. no. BB1 E02)
- Motorized linear stage (Prior, cat. no. HLD117) with focus drive (Prior, cat. no. H122)
- Oil-immersion objective (×60 total internal reflection fluorescence (TIRF), numerical aperture = 1.49; Olympus, APON 60XOTIRF)
- Optical air table (Newport)

7.1.4. Software

- Code for detection of centromere positions in 3D fluorescence microscopy images (https://github.com/TheLaueLab/blob-detection.git)
- ImageJ (https://imagej.nih.gov/ij)
- Micro-Manager (https://micro-manager.org). Software to control the microscope.
- nuc3D (in-house chromosome structure visualisation software, developed by Dr. Tim Stevens)
- NucDynamics (https://github.com/tjs23/nuc_dynamics)
- NucProcess (https://github.com/tjs23/nuc_processing)
- Python v2 (v≥2.7) or Python v3 (https://www.python.org) and the corresponding SciPy (https://www.scipy.org)/NumPy (http://www.numpy.org) modules

7.2. Single-cell Hi-C procedure

7.2.1. Cell lines and cell culture methods

For CENP-A and/or H2B imaging, stable cell lines were created by transfecting mEos3.2-tagged CENP-A plasmid and/or tandem iRFP tagged histone H2B plasmid using lipofectamine 2000, and selected by geneticin and/or puromycin respectively. Mouse ESCs were cultured on 0.2% (wt/vol) gelatin in N2B27 (NDiff B27) growth medium supplemented with 2i/LIF. Passage every 2 days, and sort for haploid cells using FACS every 4 to 8 passages as described by Freimann et al.[91]. Five to ten million haploid mESCs in a T75 flask were washed with 10 ml PBS solution, and covered with 1 ml Accutase reagent by gently tilting the flask. After incubation for 1 – 2 min at RT, or when cells were detached, single cell suspension was made by adding 10 ml of growth medium and mixing by pipetting up and down. The single cell suspension were transferred into a 15 ml centrifuge tube, and centrifuged at 500 g for 5 min. Supernatant was removed without disrupting the cell pellet.

7.2.1.1. Preparation of differentiated mESC

For single-cell Hi-C studies on differentiated mESCs, a cell line expressing a destabilised GFP protein from the endogenous *Zfp42* (*Rex1*) locus was created in haploid mESC[108]. The haploid Rex-1 GFP line was maintained and passaged in N2B27 growth medium with 2i/LIF as described in Section 7.2.1. To begin the differentiation experiment 1-2 million haploid *Rex1*-GFP cells were seeded onto 150 cm dish in N2B27 media containing 2i inhibitors (without LIF) After 24 h culture (0 h differentiated time point sample), media was replaced with fresh N2B27 media without 2i or LIF and grown for a further 24 or 48 h (24 or 48 h differentiated time point sample, respectively). At the relevant time point for each experiment (0, 24 or

48 h) cells were harvested and single cell suspensions prepared as outlined in Section 7.2.1.

7.2.2. Cell fixation and nuclear extraction

The cell pellet was resuspended in 25 ml of fresh N2B27 medium (2i and/or LIF was included if the cells had been harvested with these reagents) with 2% (vol/vol) formaldehyde in a 50 ml centrifuge tube. Cells were fixed at RT (room temperature) for exactly 10 min, with mixing by inverting the tube every 2 min. Fixation was quenched by adding 1.7 ml of 2 M glycine and mixing by gently inverting the tube. Cells were pelleted by centrifuging at 300 g for 5 min at 4℃. Supernatant was removed without disrupting the pellet. Cells were then washed in 50 ml ice-cold PBS solution, by mixing, then centrifuging at 300 g for 5 min at 4℃, and removing most of the supernatant without disrupting the pellet. The pellet was resuspended into the remaining PBS solution by gently tapping the tube. To make a control whole cell sample for later single-cell sorting using FACS, 25 μl cell suspension was added to 0.5 ml of 1× NEBuffer 3 and filtered through a 50-μm filter to remove clumps. To extract 50 ml ice-cold nuclei extraction buffer was added to the remaining cell suspension and incubated for 30 min on ice, with mixing by gently inverting the tube every 10 min. After nuclear extraction, nuclei were pelleted by centrifuging at 600 g for 5 min at 4℃. The supernatant was carefully removed without disrupting the pellet. Nuclei were washed once with 50 ml ice-cold PBS solution, by resuspending, centrifuging at 300 g for 5 min at 4℃, and removing as much of the supernatant as possible without disrupting the final nuclei pellet at the bottom of the tube. Extracted nuclei were resuspended in 1 ml 1× NEBuffer 3 and filtered through a 50-μm filter to remove clumps.

7.2.3. Single nuclei sorting by FACS

Single haploid mESC nuclei were sorted into a 384-well Cyclo-olefin polymer (COP) clear-bottom plate. Detailed setup of the FACS equipment was described in Lando et al.[66]. During the process, the FACS equipment setup aimed to pick only a single nucleus each time, sort the nucleus directly to the bottom of each well while overcoming the equipment vibration, and complete sorting all samples within 5 min to avoid nuclei drying out. To meet these requirements, single nuclei were selected according to their size by both forward and side light scattering, and only sorted ~120 wells out of 384 to minimize the effect of vibration and complete the process in time. After FACS, single nuclei were immediately humidified by adding 10 µl of 1× NEBuffer 3 to each well containing a nucleus, using a multidispenser pipette. A lid was placed on the plate to minimize evaporation.

7.2.3.1. Differentiated single nuclei sorting by FACS

Differentiated nuclei were sorted in the same way as mESCs, except an additional parameter the intensity of the *Rex1*-GFP reporter was used during FACS as follows. For the 24 h differentiated nuclei sample both the top (*Rex1*-high) and bottom (*Rex1*-low) 25% of GFP fluorescence was sorted to individual wells; whereas for the 48 h differentiated nuclei sample only cells not expressing GFP fluorescence (generally 90-95% of the total sample) were sorted.

7.2.4. 3D Imaging

To avoid nuclei drying out and altering DNA structure, imaging was carried out within 9 h of sorting single-nuclei. Detailed setup of the microscope was described in Lando et al.[66]. If cells were transfected to express fluorescent CENP-A and/or H2B (Section 7.2.1), nuclei were imaged at 488 nm for mEos3.2-tagged CENP-A, 641 nm

for tandem iRFP tagged histone H2B, and with white light; otherwise, nuclei were only imaged with white light.

During the process, the number of single nuclei was checked one after another as described below. The number of fluorescent CENP-A foci correlates closely with the number of centromeres and hence the number of chromosomes. A haploid mESC has 20 chromosomes (1 – 19 and X), but centromere regions tend to cluster together. Therefore nuclei with 10 – 20 CENP-A foci were accepted whereas the ones with less than 10 could be missing chromosomes, while those with more than 20 foci could be in S/G2-phase, diploidised or multi nuclei and were discarded. (In addition, CENP-A 3D images were also used to validate remodelled structures, see Section 7.2.12) The imaging of H2B-iRFP was used to outline the nuclear size and shape. Nuclei with no H2B signal or with unusual H2B staining were not processed further. White light imaging was used to identify the nuclei and any nuclei that did not look like normal single nuclei were discarded. In most experiments imaging was stopped when 20 single nuclei with good quality images were identified. After imaging 10μl of 1% (wt/vol) liquid LMP agarose preheated to 37℃ was gently added to the bottom of each well containing an identified nucleus. New tips were used for each well to avoid cross-contamination. With a lid on, the plate was left at RT for 30 min for the LMP agarose to solidify.

7.2.5. Single-cell in-nucleus Hi-C reactions

All steps in this section were carried out on the single-nucleus-containing the LMP agarose pad. Extra precautions were used when removing solution above the agarose pad, so as not to lose the pad or suck any agarose into the tip when pipetting, and new tips were used for each well to avoid cross-contamination. Reactions for in-nucleus Hi-C were carried out as follows.

10 μl of 0.3% SDS in 1× NEBuffer 3 was added to each well. The plate was incubated at 37℃ for 1 h. After adding 10 μl of 4% Triton X-100 in 1× NEBuffer 3 to each well,

the plate was incubated at 37℃ for another 1 h to quench the SDS. To carry out RE1 digestion, 125 U MboI in 20 μl of 1× NEBuffer 3 was then added to each well; and the plate was incubated at 37℃ for 12 – 20 h. After RE1 digestion, without disrupting the agarose pad, all solution was gently removed from each well. Biotin end-filling was carried out by adding to each well 6 μl of 1× NEB Buffer 3 containing 280 μM each of dCTP/dTTP/dGTP/biotin-14-dATP and 5 U of DNA polymerase I Large (Klenow) fragment, and incubating at 37℃ for 1 h. Then each sample was washed by adding 50 μl of 1× T4 DNA ligase buffer, leaving at RT for 10 min, and removing all solution. Finally, Hi-C junctions were ligated by adding to each well 50 μl of 1× T4 DNA ligase buffer containing 150 U of T4 DNA ligase and 5 μg of BSA, and incubating at 16℃ for 12 – 20 h.

7.2.6. Crosslink reversal

After Hi-C junction ligation, all the solution from each well was gently removed, without disrupting the agarose pad, and new tips were used for each well. Each sample was washed with 50 μl of PBS solution, by incubating at RT for 10 min and removing the solution. The plate was then sealed with a PCR adhesive plastic film. To reverse the crosslinks, the sealed plate was incubated at 65℃ for 12 – 20 h.

7.2.7. Preparing single cell Hi-C libraries for sequencing

In this section, new tips were used every time when adding and removing solutions to avoid cross-contamination. The plate was always well sealed with a PCR adhesive plastic film, except when solutions were added or removed. All mixing steps for beads and solution in plate were carried out as follows, unless specified. The 384-well plate was shaken in a plate mixer (thermomixer) at 2000 rpm for 30 seconds (s) at the temperature of the corresponding reaction. All in-plate bead washing steps were carried out as follows, unless specified. Specific washing buffer was added to each

well containing Hi-C samples bound to beads. Buffer and beads were mixed and separated with a 384-well plate magnetic separation stand for 1 min. Solution was removed by pipetting up without disrupting the beads. The residual solution was removed as possible. More than one pipetting operation was normally required to remove the solution from each well. If the beads are disrupted when pipetting up, a brown colour of the bead mix could be seen in the pipette tip. If this happens then the solution should be dispensed back into the well, separated from beads again, and sucked up again, until no brown colour is seen.

After crosslink removal, the plate was cooled down from 65℃ to 37℃, keeping the LMP agarose in solution. Alternatively, if the plate was already cooled down and stored at 4℃ and the agarose was already solid, the plate was reheated to 65℃ and incubated for 15 min to melt the agarose, and cooled down from 65℃ to 37℃. The biotinylated DNA fragments in each sample were bound to streptavidin magnetic beads by adding 20 µl 37℃ preheated streptavidin bead slurry to each well, and incubating the plate at 37℃ for 1h with mixing every 15 min in the plate mixer (thermomixer) at 37℃. During the incubation, the 384-well plate magnetic separation stand and BWT buffer (20 µl for each sample) were preheated to 37℃ for a minimum of 15 min in an incubator. After the incubation and the final mixing, while keeping the plate and magnetic stand at 37℃ in the incubator, LMP agarose was carefully removed. Because the temperature in the incubator gradually went down when pipetting with the door open, to avoid agarose solidifying in the later wells, agarose removal for all samples was carried out as fast as possible, ideally within 15 min for 20 samples. Then the beads in each well were washed with 20 µl BWT buffer in the incubator at 37℃. The beads were washed with 20 µl 1× CutSmart buffer.

Depending on the experiment, the next steps of the protocol were carried out using either the AluI-A-tailing method (6.2.7.1) or the transposase method (6.2.7.2). See Section 2.2 for the experimental design of these steps.

7.2.7.1. AluI-A-tailing method

For RE2 restriction, 20 µl 1× CutSmart buffer containing 10 U AluI was added to each well. Then the plate was incubated at 37℃ for 1 h, with mixing every 15 min. After the incubation and the final mixing, reaction solution was separated from beads and removed. Beads in each well were washed with 20 µl BWT buffer, and then 20 µl 1× NEBuffer 2. For A-tailing, 20 µl 1× NEBuffer 2 containing 0.2 mM dATP and 10 U Klenow Fragment 3'→5' exo- was added to each well. Then the plate was incubated at 37℃ for 30 min, with mixing every 10 min. After the incubation and the final mixing, reaction solution was separated from beads and removed. Beads in each well were washed with 20 µl BWT buffer, and then 20 µl 1× T4 ligase buffer. For adaptor ligation, 1 µl of 12 mM oligonucleotide adaptor with unique 3-letter barcode (different barcodes for different samples) and 19 µl 1.05× T4 ligase buffer containing 800 U T4 DNA ligase were added to each well. Then the plate was incubated at RT for 1 h, with mixing every 15 min. At this point the reaction could be stored at 4℃ overnight. After the incubations and a final mixing, reaction solution was separated from beads and removed. Beads in each well were washed twice with 20 µl BWT buffer, and then 20 µl 1× Pfx buffer.

Beads in each well were transferred to an individual 0.2-ml PCR tube, with 15 µl 1× Pfx buffer. At least two pipetting operations (10 µl + any residual solution) were carried out to ensure the transfer all ~15 µl bead slurry. For PCR library amplification, 10 µl 1× Pfx buffer containing 4 mM $MgSO_4$, 0.8 mM dNTP mix, 2 µM library amplification primer mix, and 2 U Platinum Pfx DNA polymerase was added to each tube. Bead and PCR solution were thoroughly mixed by pipetting up and down. Then the libraries were amplified with the following program: 94℃ for 2 min, 25 cycles of (94℃ for 15 s, 62℃ for 30 s and 72℃ for 1 min), 72℃ for 10 min, cool down and store at 4℃. To isolate amplified library from streptavidin beads, the PCR reaction solution was separated from beads on a magnetic separation stand. Then 20 µl supernatant for each library was carefully transferred to a 1.5 ml Eppendorf DNA Lo-bind tube without disrupting the beads. Beads were resuspended in the remaining

supernatant plus 100 µl ultrapure water, and separated from the solution again on the magnetic stand. 100 µl clear supernatant was transferred to the same 1.5 ml Lo-bind tube without disrupting the beads, ending up with 120 µl solution.

For amplified library clean-up, Agencourt AMPure XP beads (the word “bead” only in this paragraph refers to the AMPure beads not the previous streptavidin beads) were warmed to RT. 90 µl of bead slurry was added to each 120 µl solution of amplified library. Beads and solution were mixed by pipetting up and down. The mix was incubated for 5 min at RT for bead binding. Beads were separated from solution on a magnetic separation stand. Supernatant was carefully removed without disrupting the beads and discarded. With the tubes still on the magnetic stand, 200 µl 80% (vol/vol) ethanol was added to each tube, and removed after 30 s incubation. Another 200 µl 80% (vol/vol) ethanol was added, incubated and removed as much as possible. The beads were air-dried 5 – 10 min. 20 µl of 10 mM Tris-HCl pH 8.0 was added to each tube. After mixing by pipetting up and down, the bead slurry was incubated for 5 min. Beads were separated from solution on a magnetic stand. The amplified library now in the supernatant was transferred to a new Lo-bind tube.

7.2.7.2. Transposase method

For trimming, 20 µl 1× CutSmart buffer containing 1 U AluI was added to each well. Then the plate was incubated at 37℃ for 1 h, with mixing every 15 min. After the incubation and the final mixing, reaction solution was separated from beads and removed. Beads in each well were washed twice with 20 µl BWT buffer, and then twice with 20 µl Tris-HCl (pH 8.0). Samples were then processed using the Nextera XT DNA Library Preparation Kit as follows. Amplicon Tagment Mix (ATM, the transposase enzyme mix) was diluted to 1/100 original concentration in Tris-HCl pH 8.0. 2.5 µl 10 mM Tris-HCl pH 8.0, 5 µl Tagment DNA (TD) buffer and 2.5 µl of 1/100 ATM was added to each well. The beads and reagents were mixed and then incubated at 55℃ for 5 mins. The reaction was cooled down to 10℃ and 2.5 µl of

Neutralize Tagment (NT) Buffer was added to each well as soon as the temperature reached 10℃. The beads and reagents were mixed and then incubated at room temperature for 5 mins. Beads were separated from the reaction solution on a 384-well plate magnetic separation stand, and the solution was removed without disrupting the beads. Beads in each well were washed twice with 20 µl 1× BWT buffer and twice with 20 µl 10 mM Tris-HCl pH 8.0. The beads in each well were transferred from the plate with 12.5 µl of 10 mM Tris-HCl pH 8.0 to a 0.2 ml PCR tube. Then 2.5 µl of specific Nextera index i7 and 2.5 µl of i5 were added to each tube, with different combinations added to different tubes. Finally, 7.5 µl of KAPA HiFi PCR amplification mix (1× HiFi Fidelity Buffer, 0.3 µM dNTP and 0.5 U KAPA HiFi Polymerase) was added to each tube. After thoroughly mixing by pipetting up and down, the libraries were then amplified with the following program: 72℃ for 3 min, 95℃ for 30 sec, then 9 cycles of (95℃ for 10 sec, 55℃ for 30 sec, 72℃ for 30 sec), 72℃ for 5 min, cool down and store at 4℃. To isolate amplified library from streptavidin beads, the PCR reaction solution was separated from beads on a magnetic separation stand. 20 µl supernatant for each library was carefully transferred to a 1.5 ml Eppendorf DNA Lo-bind tube without disrupting the beads. Beads were resuspended in the remaining supernatant plus 100 µl ultrapure water, and separated from the solution again on the magnetic stand. 100 µl clear supernatant was transferred to the same 1.5 ml DNA Lo-bind tube without disrupting the beads, ending up with 120 µl solution. For amplified library clean-up, Agencourt AMPure XP beads were used in the same way as described in the previous section 7.2.7.1 except that libraries were eluted in a final volume of 25 ul and transferred to PCR tubes.

For the second round of PCR, 5 µl of specific Nextera index i7 and 5 µl of i5 were added to each tube, with different combinations added to different tubes. 15 µl of KAPA HiFi PCR amplification mix (1× HiFi Fidelity Buffer, 0.3 µM dNTP and 0.5 U KAPA HiFi Polymerase) was added to each tube. After thoroughly mixing by a quick vortex and a quick spin, the libraries were then amplified with the following program: 72℃ for 3 min, 95℃ for 30 sec, then 16 cycles of (95℃ for 10 sec, 55℃ for 30 sec, 72℃ for 30 sec), 72℃ for 5 min, cool down and store at 4℃. To isolate amplified

library from streptavidin beads, the PCR reaction solution was separated from beads on a magnetic separation stand. 45 µl supernatant for each library was carefully transferred to a 1.5 ml Eppendorf DNA Lo-bind tube without disrupting the beads. Beads were resuspended in the remaining supernatant plus 75 µl ultrapure water, and separated from the solution again on the magnetic stand. 75 µl clear supernatant was transferred to the same 1.5 ml DNA Lo-bind tube without disrupting the beads, ending up with 120 µl solution.

For the second round of amplified library clean-up, Agencourt AMPure XP beads were used as described in the previous section 7.2.7.1.

7.2.8. Sequencing library fragment analysis

For the remainder of the protocol the AluI-A-tailing and the transposase method were carried out using the same procedure described below, unless specified.

1 µl of each purified library was analysed on a high-sensitivity DNA chip in an Agilent 2100 Bioanalyzer, following the manufacturer's instructions. In the Bioanalyzer data visualisation software, an additional operation was set to analyse library parameters within 300 – 700 bp for mass concentration and molar concentration. For each library, mass yield within 300 – 700 bp was calculated. Libraries with over 20 ng yield within 300 – 700 bp and a centralised fragment distribution pattern between 200 – 1000 bp were selected for pooling.

7.2.9. Library pooling and size selection

Libraries generated using the same library preparation method, AluI-A-tailing or transposase, can be pooled together for a single sequencing run, whereas libraries from different methods cannot be sequenced together. As well only the libraries with different barcodes/index combinations can be pooled for a single sequencing run.

The number of moles between 300 – 700 bp per library to be pooled was calculated

based on the number of libraries and sequencing facility requirements, where each library should contribute the same number of moles. The volume from each library to be pooled was calculated using the required number of moles and the molar concentration calculated from the Bioanalyzer run (see the previous Section 7.2.8) between 300 – 700 bp. The total volume of pooled sample was summed and libraries were pooled according to the calculated volumes.

To select fragments between 300 – 700 bp from the pooled library, Agencourt AMPure XP beads were warmed to RT. Two portions of 1.8× (the volume of pooled library) of beads were transferred to two individual DNA Lo-bind tubes. Beads of both tubes were separated from the solution on a magnetic separation stand. Then 1.25× and 1.65× (the volume of pooled library) of solution was discarded from the tubes respectively. Beads were resuspended in the remaining 0.55× and 0.15× (the volume of pooled library) of solution respectively. The pooled library was added to the first 0.55× tube, mixed by pipetting up and down, and incubated at RT for 10 min. Beads were separated from solution on a magnetic stand. The supernatant was transferred to the second 0.15× tube as much as possible. After thoroughly mixing by pipetting up and down, the bead slurry in the second tube was incubated at RT for another 10 min. Beads in the second tube were separated from solution on a magnetic stand, and the supernatant was discarded. With the second tube still on the magnetic stand, 80% (vol/vol) ethanol was added to the tube enough to cover all beads on the wall of the tube, and removed after 30 s incubation. Another portion of 80% (vol/vol) ethanol enough to cover the beads was added, incubated and removed as much as possible. The beads were air-dried 5 – 10 min. 10 mM Tris-HCl pH 8.0, volume depended on sequencing facility requirements, was added to the tube. After mixing by pipetting up and down, the bead slurry was incubated for 10 min. Beads were separated from solution on a magnetic stand. The size-selected library as the supernatant was transferred to a new 1.5 ml DNA Lo-bind tube.

1 µl of the size selected library was diluted in 9 µl of 10 mM Tris-HCl pH 8.0. Then 1 µl of the 1/10 diluted library was analysed on a high-sensitivity DNA chip in an Agilent 2100 Bioanalyzer, following the manufacturer's instructions. In the

Bioanalyzer data visualisation software, an additional operation was set to analyse library parameters within 50 – 3000 bp (assumed as the whole library) for molar concentration and average fragment length. The success of size selection was checked based on the molar yield and fragment distribution pattern according to the datasheet, where fragments should be mostly found between 300 – 700 bp, the average fragment length should be within 450 – 600 bp and total molar yield (between 50 – 3000 bp) should meet sequencing facility requirements.

7.2.10. High-throughput sequencing

The size selected library was diluted if needed to meet the sequencing facility requirement on sample concentration. Then the library was sequenced using high-throughput sequencing facilities. In our cases, MiSeq paired-end 75 bp (PE75) and HiSeq PE150 were used. The MiSeq was used to check sequencing qualities of the libraries, and the good ones were pooled, size-selected and sequenced again using the HiSeq with a better sequencing depth. Two FASTQ files were generated from a paired-end sequencing run.

7.2.11. Sequencing read processing

The sequencing data in FASTQ format were processed, analysed and transformed into single-cell Hi-C specific format, which was then used to calculate 3D models of genome structure. Detailed descriptions of the software and reports were introduced in Lando et al.[66] and Stevens et al.[92]. This section briefly describes the computational pipeline.

7.2.11.1. Split barcodes

Reads in the two FASTQ files were split into individual FASTQ datasets according to the 3-letter barcodes or the transposase indexes. The two datasets with the same barcode/indexes together, each from one of the original FASTQ files, constitute data for that particular library. A python script in the NucProcess software called split_Fastq_Barcodes.py was used to carry out the split.

With the two original FASTQ files in the current directory, the command line options I used to run the script were as follows:

- python (/route_to_the_script_directory/)split_Fastq_Barcodes.py (file_name)_r_1.fq (file_name)_r_2.fq

The resultant split file names were in a format as follows:

- (file_name)_r_1_(barcode_or_indexes).fq
- (file_name)_r_2_(barcode_or_indexes).fq

7.2.11.2. NucProcess

For each pair of split FASTQ file with the same barcode/indexes, the reads were analysed and filtered in several steps to identify valid Hi-C contacts. With the two split FASTQ files in the current directory, the command line options I used to run the script were as follows:

For AluI-A-tailing libraries,

- (/route_to_the_software_directory/)nuc_process -f (/route_to_genome_index_directory/)*.fa -o (cell_name) -v –a -re1 MboI -re2 AluI -s 150-2000 -n 12 -g (/route_to_genome_build_directory/)(genome_build_name) –i (file_name)_r_?_(barcode).fq

And for transposase libraries,

- (/route_to_the_software_directory/)nuc_process -f

(/route_to_chromosome_indexing_directory/)*.fa -o (cell_name) -v –a -re1 MboI -s 150-2000 -n 12 -g (/route_to_genome_build_directory/)(genome_build_name) –i (file_name)_r_?_(indexes).fq

The command line options used have the following meaning:

- -f create genome index (e.g. according to chromosome number) and RE restriction site files
- -o output file name
- -v show processing progress in verbose output on screen
- -a generate a .ncc file containing ambiguous contacts that map to multiple sites of the genome
- -re1 MboI set RE1 as MboI
- -re2 AluI set RE2 as AluI
- -re2 not included set RE2 as random for transposase method
- -s 150-2000 valid fragment size in bp
- -n number of computer cores allowed to occupy in parallel
- -g Bowtie2 format reference genome file
- -i input files, need to be a pair of split FASTQ file

The two input FASTQ files gave several output files as follows:

- (cell_name).ncc file containing valid Hi-C contacts uniquely mapped to genome
- (cell_name)_ambig.ncc file containing valid but ambiguous Hi-C contacts
- (cell_name)_report.svg a svg image of the processing report
- (cell_name)_contact_map.svg a svg image of Hi-C contacts

7.2.11.3. NucDynamics

For libraries with enough suitable Hi-C contacts from a haploid single mouse genome, 3D models of genome structure were calculated based on the contacts using the NucDynamics software. In brief, the calculations were carried out by calculating models at low resolutions (bead sizes) first, and refining the models to higher

resolution. Models with less pairwise structural conflictions were selected at certain stages for further calculations. With the .ncc files in the current directory, from the command line, the command with options I used to run the script is as follows:

- nuc_dynamics (cell_name).ncc -m 10 -f pdb

The command line options used have the following meaning:

- -m number of models
- -f output file format

7.3. Population Hi-C procedure

7.3.1. Cell sample preparation

The cell sample was prepared in the same way as single-cell Hi-C using 30 - 40 million cells as described in Section 7.2.1. This also applies to differentiated cells described in Section 7.2.1.1.

7.3.2. Cell fixation and nuclear extraction

The cell pellet was resuspended in 40 ml of media that the cells were cultured in (2i/LIF N2B27 for ES cells, 2i N2B27 for 0 h differentiated and only N2B27 for both 24 and 48 h differentiated samples) with 1% (vol/vol) formaldehyde at RT in a 50 ml centrifuge tube. Cells were fixed for exactly 10 min, with mixing by inverting the tube every 2 min. Fixation was quenched by adding 2.5 ml of 2 M glycine and mixing by gently inverting the tube. Cells were pelleted by centrifuging at 300 g for 3 min at RT. Supernatant was removed without disrupting the pellet. Cells were then washed in 50 ml ice-cold PBS solution, by resuspending, centrifuging at 300 g for 3 min at RT, and removing supernatant without disrupting the pellet. Cells were resuspended after adding 2 ml of PBS, and filtered through a 50-μm filter to remove clumps. Cells were sorted into PBS for haploid G_1-phase nuclei by FACS. The cells were transferred to a

50 ml centrifuge tube and resuspended in 50 ml PBS. The tube was centrifuged at 300 g for 3 min at RT. Supernatant was removed without disrupting the final ~0.5 ml at the bottom. The remaining suspension was resuspended in 50 ml ice-cold nuclei extraction buffer, incubated for 30 min on ice, with mixing by gently inverting the tube every 10 min. After nuclear extraction, nuclei were pelleted by centrifuging at 600 g for 3 min at RT. The supernatant was carefully removed without disrupting the final ~0.5 ml. Nuclei were washed with 50 ml ice-cold PBS solution, by resuspending, centrifuging at 600 g for 3 min at RT, and removing supernatant without disrupting the final 1 ml.

7.3.3. In-nucleus Hi-C reactions for mES cells

Nuclei were transferred to a 1.5 ml Lo-bind tube. Supernatant was removed and nuclei were resuspended in 400 µl 1× NEBuffer 3. 12 µl of 10% SDS was added to the tube, and the tube was incubated at 37℃ for 1 h with mixing at 1000 rpm for 15 s every min. After heating, to quench SDS, 80 µl of 10% Triton X-100 was added to the tube. After mixing the tube was incubated at 37℃ for 1 h. 50 µl 1× NEBuffer 3 containing 1250 U MboI was added to the tube. After mixing, the tube was incubated at 37℃ for 12 – 20 h, with mixing at 1000 rpm for 15 s every min. End-filling was carried out by adding to the tube 50 µl of 1× NEBuffer 3 containing 280 µM each of dCTP/dTTP/dGTP/biotin-14-dATP and 50 U of DNA polymerase I Large (Klenow) fragment, and incubating at 37℃ for 1 h with mixing at 1000 rpm for 15 s every min. After end-filling, the tube was centrifuged at 700 g for 10 min at RT. Supernatant was removed and Hi-C junctions were ligated by adding to the tube 1 ml T4 ligase buffer containing 100 µg BSA and 7500 U T4 DNA ligase, and incubating at 16℃ for 12 – 20 h with mixing at 1500 rpm for 15 s every min.

7.3.3.1. In nucleus Hi-C reactions for differentiated cells

The procedure in this section is amended from Rao et al.[35] and used for population Hi-C on differentiated cells instead of the procedure described in Section 7.3.3. Nuclei were transferred to 1.5 ml Lo-bind tubes and centrifuged at 700 g for 3 min at RT. Supernatant was removed. Nuclei were gently resuspended in 50 µl of 0.5% SDS, and incubated at 62℃ for 5 min. After heating, to quench SDS, 145 µl water and 25 µl 10% Triton X-100 were added to the tube. After mixing the tube was incubated at 37℃ for 15 min. 25 µl 1× NEBuffer 2 containing 100 U MboI was added to the tube. After mixing, the tube was incubated at 37℃ for 12 – 20 h, with mixing at 1000 rpm for 15 s every min. After incubation, the tube was incubated at 62℃ for 20 min to inactivate MboI, and cooled down to RT. End-filling was carried out by adding to the tube 50 µl of water solution containing 280 µM each of dCTP/dTTP/dGTP/biotin-14-dATP and 40 U of DNA polymerase I Large (Klenow) Fragment, and incubating at 37℃ for 1 h with mixing at 1000 rpm for 15 s every min. After end-filling, Hi-C junctions were ligated by adding to the tube 900 µl ligation master mix (663 µl water, 120 µl 10× T4 ligase buffer, 100 µl 10% Triton X-100, 12 µl 10 mg/ml BSA and 5 µl 400 U/µl T4 DNA ligase), mixed by inverting the tube, and incubating at RT for 4 h with mixing by inverting the tube every 30 min. The tube was centrifuged at 600 g for 3 min at RT; supernatant was removed without disrupting the final ~50 µl/pellet. Then nuclei were resuspended in 1 ml PBS.

7.3.4. Crosslink reversal and DNA purification

The tubes for either Section 7.3.3 and/or 7.3.3.1 were centrifuged at 600 g for 3 min at RT. Supernatant was removed without disrupting the final ~50 µl/pellet. Nuclei were washed with 1 ml of PBS and resuspended in 400 µl PBS with 10 µl 20 mg/ml Proteinase K. After mixing, the tube was incubated at 65℃ for 12 – 20 h. After the incubation, another 10 µl 20 mg/ml Proteinase K and 40 µl 10% SDS was added to

the tube. The tube was then incubated at 55℃ for 30 min. 40 µl of 5M NaCl was added to the tube and the tube was cooled down to RT. 800 µl of pure ethanol and 50 µl of 3M sodium acetate was added to the tube. After mixing, the tube was incubated at -80℃ for 12 – 20 h. After incubation, the tube was centrifuged at 14000 rpm for 15 min at 4℃. Supernatant was removed without disrupting the final ~50 µl containing the DNA pellet. The DNA pellet was washed twice with 400 µl of 70% ethanol, by resuspending, centrifuging at 14000 rpm for 5 min, and removing the supernatant. The DNA pellet was briefly air dried for 1 to 2 min before resuspending in 200 µl of 10 mM Tris-HCl pH 8.0, and incubated at RT for 15 min to help dissolve the DNA. DNA concentration was measured using Nanodrop.

7.3.5. DNA shearing

Purified DNA was diluted to 10 ng/µl in 600 µl 10 mM Tris-HCl pH 8.0 in a Lo-bind tube. The diluted DNA was sheared in an ice water bath by sonication using Sonic Dismembrator with the following settings: attach small tip for 0.5 ml samples, 25% amplitude, alternation of 10 s on and 10 s off, and 3 min run time. 100 µl of sheared DNA was transferred to a new Lo-bind tube. The fragment ends were repaired by adding to the tube 12 µl 10× T4 ligase buffer, 3 µl 10 mM each dNTP, 2 µl 10 U/µl T4 PNK, 2 µl 3 U/µl T4 DNA polymerase I and 1 µl 5U/µl DNA polymerase I Large (Klenow) Fragment, mixing by gently vortex, and incubating at RT for 30 min. To purify the repaired DNA, Agencourt AMPure XP beads were warmed to RT. Then 66 µl (0.55× sample volume) of bead slurry was added to each 120 µl solution of amplified library. Beads and solution were mixed by pipetting up and down. The mix was incubated for 10 min at RT for bead binding. Beads were separated from solution on a magnetic separation stand. The supernatant was transferred to a new 1.5 ml Lo-bind tube as much as possible. Exactly 18 µl (0.15× sample volume) 4× concentrated AMPure beads were added to the transferred supernatant. After thoroughly mixing by pipetting up and down, the bead slurry in the second tube was

incubated at RT for another 10 min. Beads in the second tube were separated from solution on a magnetic stand, and the supernatant was discarded. With the second tube still on the magnetic stand, 200 µl 80% (vol/vol) ethanol was added to the tube, and removed after 30 s incubation. Another portion of 200 µl 80% (vol/vol) ethanol was added, incubated and removed as much as possible. The beads were air-dried 5 – 10 min. Then 50 µl 10 mM Tris-HCl pH 8.0 was added to the tube. After mixing by pipetting up and down, the bead slurry was incubated for 10 min. Beads were separated from solution on a magnetic stand. Supernatant was transferred to a new 1.5 ml DNA Lo-bind tube. 10 µl of DNA solution after shearing and 10 µl of DNA solution after purification were analysed on a 2% agarose gel to verify successful size selection.

7.3.6. Sequencing library preparation using the A-tailing method

40 µl Dynabeads M-280 streptavidin was transferred to a 1.5 ml tube. Beads were separated from solution on a magnetic separation stand and supernatant was removed. The beads were washed with 40 µl 2× BW buffer and resuspended in 40 µl 2× BW buffer. The 40 µl streptavidin beads slurry was added to the 40 µl size selected DNA. After mixing by gentle vortex, the tube was incubated at RT for 30 min in the Thermomixer, with mixing at 1500 rpm for 15 s every min and gentle vortex every 10 min. After binding, the beads were separated from solution on a magnetic stand, and supernatant was removed without disrupting the beads. The beads were then washed with twice with 100 µl 1× BWT buffer and twice with 100 µl 1× NEBuffer 2. For A-tailing, the beads were resuspended in 50 µl 1× NEBuffer 2 containing 0.2 mM dATP, 5 U Klenow fragment 3’ – 5’ exo-, and incubated at 37℃ for 30 min in the thermomixer, with mixing at 1500 rpm for 15 s every min. After A-tailing, the reaction solution was separated from beads and removed and the beads were washed with 100 µl 1× BWT buffer and twice with 100 µl 1× T4 ligase buffer. For adaptor ligation, 2.5 µl of 12 mM oligonucleotide adaptor with selected 3-letter barcode and

47.5 µl 1.05× T4 ligase buffer containing 800 U T4 DNA ligase were added to the tube. Then the tube was incubated at RT for 1 h in the thermomixer, with mixing at 1500 rpm for 15 s every min. Then the reaction might be stored at 4℃ overnight. After the incubations and a final mixing, reaction solution was separated from beads and removed. Beads were washed twice with 100 µl BWT buffer and then twice with 100 µl 1× Pfx buffer, resuspended in 40 µl 1× Pfx buffer, and equally transferred (20 µl × 2) to two 0.2 ml PCR tubes. The 20 µl 1× Pfx buffer was separated from beads and removed.

For PCR library amplification, 25 µl 1× Pfx buffer containing 2 mM $MgSO_4$, 0.4 mM dNTP mix, 1 µM library amplification primer mix, and 2 U Platinum Pfx DNA polymerase was added to each tube. Bead and PCR mix were thoroughly mixed by pipetting up and down. Then the libraries were amplified with the following program: 94℃ for 2 min, 10 cycles of (94℃ for 15 s, 62℃ for 30 s and 72℃ for 1 min), 72℃ for 10 min, cool down and store at 4℃. To isolate amplified library from streptavidin beads, the PCR reaction solution was separated from beads on a magnetic separation stand. 20 µl supernatant from the two libraries was carefully transferred to the same new 1.5 ml Eppendorf DNA Lo-bind tube without disrupting the beads. Beads in each tube were resuspended in the remaining supernatant plus 100 µl ultrapure water, and separated from the solution again on the magnetic stand. 100 µl clear supernatant was transferred to the same 1.5 ml Lo-bind tube without disrupting the beads, ending up with 240 µl solution.

For amplified library clean-up and size selection, Agencourt AMPure XP beads (the word “bead” only in this paragraph refers to this AMPure beads not the previous streptavidin beads) were warmed to RT. 144 µl (0.6× sample volume) of bead slurry was added to the 240 µl solution of amplified library. Beads and solution were mixed by pipetting up and down. The mix was incubated for 10 min at RT for bead binding. Beads were separated from solution on a magnetic separation stand. The supernatant was transferred to a new 1.5 ml Lo-bind tube as much as possible. Exactly 24 µl (0.1× sample volume) 4× concentrated AMPure beads were added to the transferred supernatant. After thoroughly mixing by pipetting up and down, the bead slurry in the

second tube was incubated at RT for another 10 min. Beads in the second tube were separated from solution on a magnetic stand, and the supernatant was discarded. With the second tube still on the magnetic stand, 500 µl 80% (vol/vol) ethanol was added to the tube, and removed after 30 s incubation. Another portion of 500 µl 80% (vol/vol) ethanol was added, incubated and removed as much as possible. The beads were air-dried 5 – 10 min. 50 µl 10 mM Tris-HCl pH 8.0 was added to the tube. After mixing by pipetting up and down, the bead slurry was incubated for 10 min. Beads were separated from solution on a magnetic stand. The supernatant containing the purified and size-selected library was transferred to a new 1.5 ml DNA Lo-bind tube.

7.3.7. Sequencing library fragment analysis

1 µl of library was analysed on a high-sensitivity DNA chip in an Agilent 2100 Bioanalyzer, following the manufacturer's instructions. In the Bioanalyzer data visualisation software, two additional operations were set to analyse library parameters within 300 – 700 bp for mass concentration and molar concentration, and within 50 – 3000 bp for molar concentration and average fragment length. A good population sequencing library should have fragments concentrated within 300 – 650 bp range, and a yield enough for sequencing.

7.3.8. High-throughput sequencing

The library was diluted if needed to meet sequencing facility requirement on sample concentration. Hi-C libraries processed using the AluI-A-tailing method, no matter single-cell or population can be pooled together for one sequencing run. However, population Hi-C libraries require much more sequencing depth than single-cell libraries. As an example, for the 400 M read capacity of a HiSeq 4000 paired-end (PE) 150bp sequencing run, a population Hi-C library normally needs at least 40% (160 M) capacity whereas a single-cell library normally needs about 2% (8 M). Then the

library was sequenced on HiSeq 2000 or 4000 with PE150 reads. Two FASTQ files were generated from a paired-end sequencing run.

7.3.9. Sequencing read processing

The sequencing data in FASTQ format were processed, analysed and transformed into Hi-C specific format, which was then used to calculate A/B compartments. Detailed descriptions of the software and reports were described in Lando et al.[66] and Stevens et al.[92]. This section briefly introduces the computational pipeline.

7.3.9.1. Split barcodes

This step for population Hi-C sample is the same as the step for single-cell Hi-C, as described in Section 7.2.11.1.

7.3.9.2. NucProcess

For each pair of split FASTQ file with the same barcode/indexes, the reads were analysed and filtered in several steps to identify valid Hi-C contacts. With the two split FASTQ files in the current directory, from the command line, the command with options I used to run the script is as follows:

- (/route_to_the_software_directory/)nuc_process -f (/route_to_genome_index_directory/)*.fa -o (cell_name) -v -p -re1 MboI -s 150-2000 -n 12 -g (/route_to_genome_build_directory/)(genome_build_name) –i (file_name)_r_?_(barcode).fq

The command line options used have meaning as follows:

- -f create genome index (e.g. according to chromosome number) and RE restriction site files

- -o output file name
- -v show processing progress in verbose output on screen
- -p indicate the input files as a population Hi-C sample
- -re1 MboI set RE1 as MboI
- -s 150-2000 valid fragment size in bp
- -n number of computer cores allowed to occupy in parallel
- -g Bowtie2 format reference genome file
- -i input files, need to be a pair of split FASTQ file

The two FASTQ files were combined to give several files as follows:

- (cell_name).ncc file containing valid Hi-C contacts uniquely mapped to genome
- (cell_name)_report.svg a svg image of the processing report
- (cell_name)_contact_map.svg a svg image of Hi-C contacts

7.3.9.3. A/B compartment calculation

The A/B compartment calculation was carried out as described in Stevens et al.[92], which was similar to the original way described in the first Hi-C paper by Lieberman-Aiden et al.[17].

7.4. Procedures for flatness analysis by moment of inertia

Data of calculated genome structure models at 100 kb resolution (see section 7.2.11.3) were transformed to 3D coordinates of unit length, where each point represented a bin of 100,000 DNA base pairs. By signing each point a unit mass, the moment of inertia (*I*) was calculated on three pseudo-axes of the three dimensions, for all individual chromosomes and the whole genome of each model. The *I* values were then averaged from 10 models of each single cell structure. *I* ratios ((sqrt(I_y/I_x) and sqrt(I_z/I_x)) were

calculated using the averaged *I* values, for all chromosomes and the whole genome of each cell.

I ratios of each chromosome (1 – 19, X) or the whole genome from all the cells at each time point and condition (involving ES, 24 h *Rex1*^{low}, 24 h *Rex1*^{high} and 48 h) form a dataset. Each dataset was plotted in boxplots to show its distribution. The normality of the distribution was tested using probability plot for normal distribution and its correlation coefficient at 1% significant level[100,101]. For each chromosome or genome, datasets from all four time points and conditions were grouped. The group variance was tested for equality by comparing the ratio of largest to smallest variance. Based on the results of normality and equal variance tests, a type of statistical test was chosen to compare the datasets within each group. Instead of setting a fixed significance level, p-values from the tests were used to comprehend the results.

7.5. Other procedures

Other procedures of single-cell Hi-C experiment and complementary experiments were carried out by other members of my group or collaborators in other groups. These include structure validation using CENP-A images, ChIP-Seq and RNA-Seq experiments. Details of these procedures can be found in Stevens et al.[92] and Lando et al.[66].

References

1. Szalaj, P. & Plewczynski, D. Three-dimensional organization and dynamics of the genome. *Cell Biol. Toxicol.* 1–24 (2018). doi:10.1007/s10565-018-9428-y
2. Roy, S. S., Mukherjee, A. K. & Chowdhury, S. Insights about genome function from spatial organization of the genome. *Hum. Genomics* **12**, 8 (2018).
3. Watson, J. & Crick, F. Molecular structure of nucleic acids. *Nature.* **171**, 737–8 (1953).
4. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
5. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* (2011). doi:10.1038/cr.2011.22
6. Robinson, P. J. J., Fairall, L., Huynh, V. A. T. & Rhodes, D. EM measurements define the dimensions of the '30-nm' chromatin fiber: Evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci.* **103**, 6506–6511 (2006).
7. Bednar, J. *et al.* Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc. Natl. Acad. Sci.* **95**, 14173–14178 (1998).
8. Van Holde, K. & Zlatanova, J. Chromatin higher order structure: Chasing a mirage? *Journal of Biological Chemistry* **270**, 8373–8376 (1995).
9. Joti, Y. *et al.* Chromosomes without a 30-nm chromatin fiber. *Nucleus* **3**, 404–410 (2012).
10. Flemming, W. Zellsubstanz, kern und zelltheilung. *F.C.W. Vogel, Leipzig*, 419 (1882).
11. Cremer, T. *et al.* Rabl's model of the interphase chromosome arrangement tested in Chinise hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Hum. Genet.* **60**, 46–56 (1982).
12. Zorn, C., Cremer, C., Cremer, T. & Zimmer, J. Unscheduled DNA synthesis

after partial UV irradiation of the cell nucleus. Distribution in interphase and metaphase. *Exp. Cell Res.* **124**, 111–119 (1979).

13. Rappold, G. A. *et al.* Sex chromosome positions in human interphase nuclei as studied by in situ hybridization with chromosome specific DNA probes. *Hum. Genet.* **67**, 317–325 (1984).
14. Manuelidis, L. Individual interphase chromosome domains revealed by in situ hybridization. *Hum. Genet.* **71**, 288–93 (1985).
15. Schardin, M., Cremer, T., Hager, H. D. & Lang, M. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum. Genet.* **71**, 281–287 (1985).
16. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, 0826–0842 (2005).
17. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
18. Branco, M. R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4**, 780–788 (2006).
19. Boyle, S. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211–219 (2001).
20. Tanabe, H. *et al.* Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc. Natl. Acad. Sci.* **99**, 4424–4429 (2002).
21. Neusser, M., Schubel, V., Koch, A., Cremer, T. & Müller, S. Evolutionarily conserved, cell type and species-specific higher order chromatin arrangements in interphase nuclei of primates. *Chromosoma* **116**, 307–320 (2007).
22. Mayer, R. *et al.* Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. *BMC Cell Biol.* **6**, (2005).
23. Koehler, D. *et al.* Changes of higher order chromatin arrangements during major genome activation in bovine preimplantation embryos. *Exp. Cell Res.* **315**, 2053–2063 (2009).

24. Kozubek, S. *et al.* 3D Structure of the human genome: order in randomness. *Chromosoma* **111**, 321–331 (2002).
25. Cremer, M. *et al.* Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei. *J. Cell Biol.* **162**, 809–820 (2003).
26. Murmann, A. E. *et al.* Local gene density predicts the spatial position of genetic loci in the interphase nucleus. *Exp. Cell Res.* **311**, 14–26 (2005).
27. Goetze, S. *et al.* The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. *Mol. Cell. Biol.* **27**, 4475–87 (2007).
28. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harbor perspectives in biology* **2**, (2010).
29. Parada, L. A., McQueen, P. G. & Misteli, T. Tissue-specific spatial organization of genomes. *Genome Biol.* (2004).
30. Walter, J., Schermelleh, L., Cremer, M., Tashiro, S. & Cremer, T. Chromosome order in HeLa cells changes during mitosis and early G1, but is stably maintained during subsequent interphase stages. *J. Cell Biol.* **160**, 685–697 (2003).
31. Cvačková, Z., Mašata, M., Staněk, D., Fidlerová, H. & Raška, I. Chromatin position in human HepG2 cells: Although being non-random, significantly changed in daughter cells. *J. Struct. Biol.* **165**, 107–117 (2009).
32. Chadwick, B. P. & Willard, H. F. Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome. *Proc. Natl. Acad. Sci.* **101**, 17450–17455 (2004).
33. Solovei, I. *et al.* Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. *Cell* **137**, 356–368 (2009).
34. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).
35. Rao, S. S. P. P. *et al.* A 3D map of the human genome at kilobase resolution

reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

36. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
37. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
38. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
39. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
40. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110–1121 (2016).
41. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112**, E6456–E6465 (2015).
42. Doyle, B., Fudenberg, G., Imakaev, M. & Mirny, L. A. Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions. *PLoS Comput. Biol.* **10**, (2014).
43. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. in *Nature Genetics* **43**, 630–638 (2011).
44. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* (2016). doi:10.1016/j.celrep.2016.04.085
45. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).
46. Vietri Rudan, M. *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep.* **10**, 1297–1309 (2015).
47. Hsieh, T. H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108–119 (2015).
48. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**,

335–348 (2012).

49. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).

50. Lin, Y. C. *et al.* Global changes in the nuclear positioning of genes and intra-and interdomain genomic interactions that orchestrate B cell fate. *Nat. Immunol.* **13**, 1196–1204 (2012).

51. Fasulo, B. *et al.* The Drosophila Mi-2 Chromatin-Remodeling Factor Regulates Higher-Order Chromatin Structure and Cohesin Dynamics In Vivo. *PLoS Genet.* **8**, (2012).

52. Qiu, Z. *et al.* Functional Interactions between NURF and Ctcf Regulate Gene Expression. *Mol. Cell. Biol.* **35**, 224–237 (2015).

53. Nikalayevich, E. & Ohkura, H. The NuRD nucleosome remodelling complex and the NHK-1 kinase are required for chromosome condensation in oocytes. **7094**, 566–575 (2015).

54. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

55. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).

56. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).

57. Bohn, M. *et al.* Localization microscopy reveals expression-dependent parameters of chromatin nanostructure. *Biophys. J.* (2010). doi:10.1016/j.bpj.2010.05.043

58. Matsuda, A. *et al.* Condensed mitotic chromosome structure at nanometer resolution using PALM and EGFP- histones. *PLoS One* (2010). doi:10.1371/journal.pone.0012768

59. Ricci, M. A., Manzo, C., Garc ́a-Parajo, M. F., Lakadamyali, M. & Cosma, M. P. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell* (2015). doi:10.1016/j.cell.2015.01.054

60. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-.).* (2007). doi:10.1126/science.1141319

61. Cockburn, K. & Rossant, J. Making the blastocyst: Lessons from the mouse. *Journal of Clinical Investigation* (2010). doi:10.1172/JCI41229

62. Boroviak, T. & Nichols, J. The birth of embryonic pluripotency. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2014). doi:10.1098/rstb.2013.0541

63. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* (2008). doi:10.1038/nature06968

64. Ying, Q. L., Nichols, J., Chambers, I. & Smith, A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* (2003). doi:10.1016/S0092-8674(03)00847-X

65. Leeb, M. & Wutz, A. Derivation of haploid embryonic stem cells from mouse embryos. *Nature* (2011). doi:10.1038/nature10448

66. Lando, D. *et al.* Combining fluorescence imaging with Hi-C to study 3D genome architecture of the same single cell. *Nat. Protoc.* **13**, 1034–1061 (2018).

67. Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).

68. Terryn, J., Tricot, T., Gajjar, M. & Verfaillie, C. Recent advances in lineage differentiation from stem cells: hurdles and opportunities? *F1000Research* **7**, 220 (2018).

69. Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev.* (1998). doi:10.1101/gad.12.13.2048

70. Kalkan, T. *et al.* Tracking the embryonic stem cell transition from ground state pluripotency. *Development* (2017). doi:10.1242/dev.142711

71. Shi, W. *et al.* Regulation of the Pluripotency Marker Rex-1 by Nanog and Sox2.

J. Biol. Chem. (2006). doi:10.1074/jbc.M601811200

72. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* (2006). doi:10.1038/nature05284
73. Adams, C. D., Schnurr, B., Marko, J. F. & Reznikoff, W. S. Pulling Apart Catalytically Active Tn5 Synaptic Complexes Using Magnetic Tweezers. *J. Mol. Biol.* (2007). doi:10.1016/j.jmb.2006.12.064
74. Reznikoff, W. S. *et al.* Tn5: A molecular window on transposition. *Biochem. Biophys. Res. Commun.* (1999). doi:10.1006/bbrc.1999.1891
75. Davies, D. R., Goryshin, I. Y., Reznikoff, W. S. & Rayment, I. Three-dimensional structure of the tn5 synaptic complex transposition intermediate. *Science (80-.).* (2000). doi:10.1126/science.289.5476.77
76. Goryshin, I. Y. & Reznikoff, W. S. Tn5 in vitro transposition. *J. Biol. Chem.* (1998). doi:10.1074/jbc.273.13.7367
77. Reznikoff, W. S. Tn5 transposition: a molecular tool for studying protein structure–function. *Biochem. Soc. Trans.* (2006). doi:10.1042/bst20060320
78. Gradman, R. J. & Reznikoff, W. S. Tn5 synaptic complex formation: Role of transposase residue W450. *J. Bacteriol.* (2008). doi:10.1128/JB.01488-07
79. Weinreich, M. D., Gasch, A. & Reznikoff, W. S. Evidence that the cis preference of the Tn5 transposase is caused by nonproductive multimerization. *Genes Dev.* (1994). doi:10.1101/gad.8.19.2363
80. Zhou, M. & Reznikoff, W. S. Tn5 transposase mutants that alter DNA binding specificity. *J. Mol. Biol.* (1997). doi:10.1006/jmbi.1997.1188
81. Steiniger, M., Metzler, J. & Reznikoff, W. S. Mutation of Tn5 transposase β-loop residues affects all steps of Tn5 transposition: The role of conformational changes in Tn5 transposition. *Biochemistry* (2006). doi:10.1021/bi061227v
82. Vaezeslami, S., Sterling, R. & Reznikoff, W. S. Site-directed mutagenesis studies of Tn5 transposase residues involved in synaptic complex formation. *J. Bacteriol.* (2007). doi:10.1128/JB.00524-07
83. Naumann, T. A. & Reznikoff, W. S. Trans catalysis in Tn5 transposition. *Proc.*

Natl. Acad. Sci. U. S. A. **97**, 8944–9 (2000).

84. Reznikoff, W. S. Tn5 as a model for understanding dna transposition. *Molecular Microbiology* (2003). doi:10.1046/j.1365-2958.2003.03382.x

85. Syed, F., Grunenwald, H. & Caruccio, N. Next-generation sequencing library preparation: Simultaneous fragmentation and tagging using in vitro transposition. *Nat. Methods* (2009). doi:10.1038/nmeth1109-802

86. Syed, F., Grunenwald, H. & Caruccio, N. Optimized library preparation method for next-generation sequencing. *Nat. Methods* **6**, i–ii (2009).

87. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).

88. Parkinson, N. J. *et al.* Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res.* **22**, 125–133 (2012).

89. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively-scaled sequencing projects. *Genome Res.* gr.177881.114- (2014). doi:10.1101/gr.177881.114

90. Illumina. Nextera XT DNA Library Preparation Guide - 15031942 - D. (2014).

91. Freimann, R., Kramer, S., Böhmler, A. & Wutz, A. Stammzellen: Gewinnung haploider Stammzellkulturen der Maus für genetische Screens. *BioSpektrum* (2014). doi:10.1007/s12268-014-0458-6

92. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).

93. Adey, A. *et al.* In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).

94. Lando, D., Stevens, T. J., Basu, S. & Laue, E. D. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: An evaluation of single-cell Hi-C protocols. *Nucleus* **1034**, 00–00 (2018).

95. Kuznetsova, I., Podgornaya, O. & Ferguson-Smith, M. A. High-resolution organization of mouse centromeric and pericentromeric DNA. *Cytogenet.*

Genome Res. **112**, 248–255 (2006).

96. Garagna, S., Zuccotti, M., Capanna, E. & Redi, C. A. High-resolution organization of mouse telomeric and pericentromeric DNA. *Cytogenet. Genome Res.* **96**, 125–129 (2002).
97. Wang, S. *et al.* Spatial organization of chromatin domains and compartments in single chromosomes. *Science (80-.).* (2016). doi:10.1126/science.aaf8084
98. Peric-Hupkes, D. *et al.* Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. *Mol. Cell* (2010). doi:10.1016/j.molcel.2010.03.016
99. Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res.* (2013). doi:10.1101/gr.141028.112
100. Filliben, J. J. The probability plot correlation coefficient test for normality. *Technometrics* (1975). doi:10.1080/00401706.1975.10489279
101. Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. a. *Notched Box Plots Graphical Methods for Data Analysis Chambers_1983.pdf. Graphical methods for data analysis* (1983). doi:10.3197/096327109x404771
102. Morgan, C. J. Use of proper statistical techniques for research studies with small samples. *Am. J. Physiol. Cell. Mol. Physiol.* (2017). doi:10.1152/ajplung.00238.2017
103. Box, G. E. P. Non-Normality and Tests on Variances. *Biometrika* (2006). doi:10.2307/2333350
104. Nagano, T. *et al.* Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat. Protoc.* **10**, 1986–2003 (2015).
105. Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).
106. Chan, E. M. *et al.* Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Cell* (2016). doi:10.1093/nar/gkt1348

107. Palayret, M. *et al.* Virtual-'light-sheet' single-molecule localisation microscopy enables quantitative optical sectioning for super-resolution imaging. *PLoS One* (2015). doi:10.1371/journal.pone.0125438

108. Leeb, M., Dietmann, S., Paramor, M., Niwa, H. & Smith, A. Genetic exploration of the exit from self-renewal using haploid embryonic stem cells. *Cell Stem Cell* (2014). doi:10.1016/j.stem.2013.12.008